Signal Processing 108 (2015) 136–146

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Whole-body humanoid robot imitation with pose similarity evaluation

Jie Lei [a], Mingli Song [a,*], Ze-Nian Li [b], Chun Chen [a]

[a] Zhejiang University, Hangzhou 310027, PR China
[b] Simon Fraser University, Burnaby, BC, Canada, V5A1S6

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form
24 July 2014
Accepted 17 August 2014
Available online 26 August 2014

Keywords:
Humanoid robot
Pose transfer
Imitation
Similarity metric

ABSTRACT

Imitation is considered to be a kind of social learning that allows the transfer of information, actions, behaviors, etc. Whereas current robots are unable to perform as many tasks as human, it is a natural way for them to learn by imitations, just as human does. With the humanoid robots being more intelligent, the field of robot imitation has getting noticeable advance.

In this paper, we focus on the pose imitation between a human and a humanoid robot and learning a similarity metric between human pose and robot pose. In contrast to recent approaches that capture human data using expensive motion captures or only imitate the upper body movements, our framework adopts a Kinect instead and can deal with complex, whole body motions by keeping both single pose balance and pose sequence balance. Meanwhile, different from previous work that employs subjective evaluation, we propose a pose similarity metric based on the shared structure of the motion spaces of human and robot. The qualitative and quantitative experimental results demonstrate a satisfactory imitation performance and indicate that the proposed pose similarity metric is discriminative.

1. Introduction

With the development of robotics, robots are getting much smarter than they used to, especially for humanoid robots. However, they are not ready to perform many tasks as naturally as human beings. Imitation is considered as an effective solution to the problem. Specifically, imitation is an advanced behavior whereby an individual observes and replicates the behaviors of others. Robots have replaced humans in the assistance of performing repetitive and dangerous tasks in some fields, such as construction industry, medical surgery, toxic substances cleaning and space exploration, where they can take advantage of imitating human to some degree.

Imitation is about generating stable humanoid movements from the human motions, an overview and computational approaches to this problem can be found in [1]. Many of the imitation researches focus on the upper body and employ complex system setting. In [2], an analytical method was proposed to transfer the upper body motion from human to humanoid robot. Riley et al. [3] made use of some colored marks on a human upper body in order to be abstracted by a vision system based on external cameras and a head-mounted one of a humanoid. These marks were used to estimate the angular range of some joints with a kinematic model of the human to perform the imitation. Similar to [3], with the help of 34 markers placed on the human upper body and 2 markers attached on a conductor stick, Ott et al. [4] applied the data obtained by a motion capture system to allow a humanoid

* Corresponding author.
E-mail addresses: ljaylei@zju.edu.cn (J. Lei), brooksong@zju.edu.cn (M. Song), li@sfu.ca (Z.-N. Li), chenc@zju.edu.cn (C. Chen).

http://dx.doi.org/10.1016/j.sigpro.2014.08.030
0165-1684/

robot to mimic the human motion regarding to a Cartesian control approach. Aleotti et al. [5] adopted neural networks to learn a mapping between the positions of a human arm and an industrial robot arm. Based on Aleotti's work, Stanton et al. [6] extended it to a humanoid robot by training a feed forward neural network with particle swarm optimization for each degree of freedom (DOF). In the data collecting process, a robot was used to lead a human operator through series of paired synchronized movements captured by a motion capture, which was time-consuming and tedious. As they were mentioned, in order to ensure robot stability, the position of the robot's ankles did not employ neural networks. Since the neural networks could not always output ideal angles, the robot, as a rigid body, was apt to lose its balance. Meanwhile, a unified neural network training for the whole body was infeasible, considering convergence trouble. Whereas training with separate networks would cause correlation loss among the DOFs. Other imitation researches are mainly dedicated to humanoid gait or walking movements [7–9]. In conclusion, existing works have the following limitations:

- Imitation of the upper body or a single part is insufficient to meet the needs of humanoid robot [2–5].
- With requiring motion capture equipment, it is expensive and inconvenient for general use and unnatural for human–robot interaction [5,6,10].
- Lack of balance control and the whole body control [3,4,6].
- The imitation results are not qualitatively evaluated [6,10,11].

After performing the pose imitation, another important issue is "how can we evaluate the imitation similarity between a robot slave and the master". In [11], Zuher et al. gave a subjective evaluation by taking persons to mark the quality of an imitation with bad, poor, fair, good and excellent. Other existing research efforts are basically concentrated on the pose similarity of a single agent. The simplest metric is $L2$ distance, which does not sufficiently utilize the data dependency between DOFs. In [12], different weights were learned for DOFs, in correspondence with the fact that some DOFs had more influences on determining the similarity. Chen et al. [13] proposed a new rich pose feature set to effectively encode the pose similarity by utilizing features on geometric relations among body parts. Based on the pose feature set, a distance metric was learned in a semi-supervised manner. By matching the related DOFs of a robot and a human, we can apply these methods to evaluate the imitation similarity. However, robots and humans are different in DOF dimensions and physical constrains, i.e., they have different motion spaces. It is inappropriate to compare them directly.

The problem we are facing here can be regarded as a metric learning problem. Learning a good distance metric in feature space is crucial in real-world applications. Good distance metrics are important to many computer vision tasks, such as image classification [14–16], content-based image retrieval [17,18] and their applications [19,20]. Many useful algorithms and ideas were proposed in these papers to combine multiple feature sets, such as high-order distance-based multiview stochastic learning (HD-MSL [14]) and semi-supervised multiview distance metric learning (SSM-DML [16]). In our case, we believe that the human poses and the humanoid robot poses have much in common for their highly similar skeleton structures. Their differences depend on the number of DOFs and physical constrains. As a consequence, the shared motion space between the two agents can be a good metric space to study the pose similarity.

This paper proposes a novel humanoid robot imitation framework with pose similarity metric learning between human pose and robot pose, using a consumer camera (the Microsoft Kinect) and a humanoid robot (the Aldebaran Nao H25). The proposed framework summarized in Fig. 1 adopts dynamic balance control with realtime imitation performance. A shared representation of both robot pose and human pose is learned to evaluate the imitation similarity. Both qualitative and quantitative experimental results demonstrate a satisfactory imitation performance and indicate that the proposed pose similarity evaluation is discriminative.

Our main contributions are the following: (a) we propose a novel framework to perform pose imitation on the whole body motions rather than the upper body. (b) We actively keep single pose balance and introduce transient poses to achieve smooth pose sequence balance. (c) We demonstrate how shared structure can provide a quantitative evaluation to define the similarity between a human pose and a robot pose.

2. Humanoid robot imitation

2.1. Pose representation

The Kinect consists of a RGB camera, a depth sensor and provides 3D human skeleton tracking at 30 frames per second. Based on the position data obtained, we can calculate 20 DOF angles listed in Table 1, which are angles between pairs of related vectors. For example,

$$\theta^{H}_{HeadPitch} = \langle DV(Pos_{Spine}, Pos_{ShoulderCenter}), DV(Pos_{ShoulderCenter}, Pos_{Head})\rangle \quad (1)$$

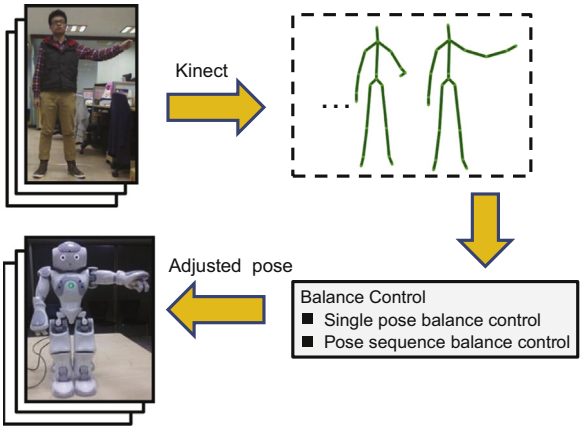

Fig. 1. Overview of our human to humanoid robot imitation system.

Table 1
List of DOFs. The DOFs with a star symbol belong to the Nao robot pose only. Both the human pose and the Nao robot pose have the rest 20 DOFs.

Body Part	DOFs
Head	HeadYaw, HeadPitch
Left Arm	LShoulderPitch, LShoulderRoll, LElbowYaw, LElbowRoll, LWristYaw*
Right Arm	RShoulderPitch, RShoulderRoll, RElbowYaw, RElbowRoll, RWristYaw*
Left Leg	LHipPitch, LHipRoll, LKneePitch, LAnklePitch, LAnkleRoll
Right Leg	RHipPitch, RHipRoll, RKneePitch, RAnklePitch, RAnkleRoll
Hip	LHipYawPitch*, RHipYawPitch*

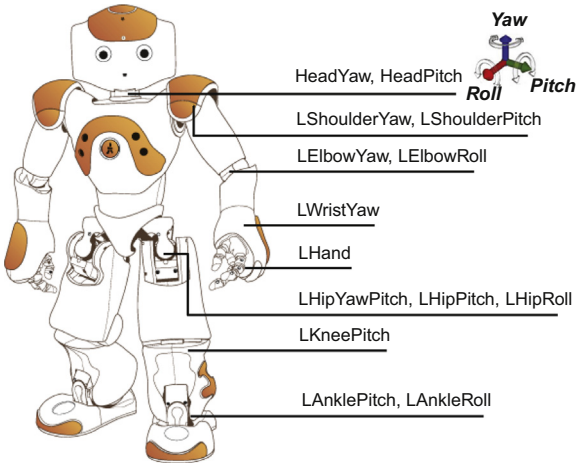

Fig. 2. The DOFs of a Nao robot, showing the head and the left body part only.

where *DV* stands for the direction vector of two 3D points. Then we get 20 angles in total for a skeleton (or a human pose), denoted as $\boldsymbol{\theta}^H = \{\theta_d^H\}$ where d is the name of a DOF.

The Nao robot owns 26 DOFs (Fig. 2), where we exclude the "LHand" and "RHand" DOFs and choose the rest 24 DOFs, as listed in Table 1, to represent a Nao pose, denoted as $\boldsymbol{\theta}^N = \{\theta_d^N\}$. As we transfer different DOF configurations, the Nao robot can display varied poses.

We choose DOF angles instead of position data to represent a pose mainly for two reasons:

- The output values of a Kinect are specified in relation to its origin coordinates in 3D space, and are easily affected to different human agents.
- Since a Nao robot is equipped with position sensors only on the endpoints of the limbs, the robot is easier to be driven by DOF angles rather than position data.

2.2. Support leg

Balance is an important issue to be considered when preforming imitation on humanoid robot. At every point in time, we need to ensure the robot is in a statically stable configuration. Specifically, the ground project of the center of mass (COM) should lie within the convex hull of the foot contact points (or support polygon for short) [21].

The support leg should be figured out before controlling the balance. There are three situations, i.e. *LLeg, RLeg,* and *Legs*, short for the left leg, the right leg and both legs, respectively. Since the positions of both feet in a human pose are known from the Kinect, we have

$$SL = \begin{cases} Legs & : |Pos^Y_{FootLeft} - Pos^Y_{FootRight}| \le \lambda \\ LLeg & : Pos^Y_{FootLeft} - Pos^Y_{FootRight} < \lambda \\ RLeg & : Pos^Y_{FootLeft} - Pos^Y_{FootRight} > \lambda \end{cases} \quad (2)$$

where *SL* is short for support leg and λ is a threshold for smoothing.

Due to the noise in Kinect data, a single threshold will fail in some cases. To deal with the data noise, a fixed-size queue is kept in advance to record the *SL*s of a small sequence starts at the pose to be imitated, then we scan the queue to find outliers and update them according to the *SL* before and after them. After imitating the first pose, we read in the next one and move the queue forward.

2.3. Transient poses

In the experiment, we noted that the Nao robot was unable to transfer from some poses to others directly in a safe way, while the human can achieve that easily, especially when the two poses were not supported by the same leg. This is because the physical constrains of humans and Nao robots are different. To solve this problem, we introduce three transient poses, as shown in Fig. 3. By inserting the transient poses into the original sequence where the adjacent poses do not belong to the same supporting case, a stable pose transfer can be achieved.

The "StandInit" pose is a built-in pose in the Nao robot system and can be regarded as a very stable pose for the two legs supporting situation. The "LeftLeaning" pose is obtained by transferring weight from the center to the left leg and can be considered as the critical pose between *Legs* supporting and *LLeg* supporting. Likewise, the "RightLeaning" pose can be considered as the critical pose between *Legs* and *RLeg*. All the three transient poses are prestored and can be transferred to the Nao robot when needed.

2.4. Balance control

To avoid falls of the robot that might occur when using direct imitation of the joint angle trajectories due to the different weight distributions, we developed a strategy to actively balance the COM. Our balance control is constituted of two parts, the single pose balance control and the pose sequence balance control.

The single pose balance control aims at adjusting a human pose to the robot pose according to the balance rule. Given a human pose $\boldsymbol{\theta}^H$, we choose the 20 corresponding angles and set the rest four angles as 0 to form a target pose of the Nao robot. Denote $\boldsymbol{X}_{COM}$ as the COM position of the robot, $\boldsymbol{\theta}_r$ as the current angle vector, i.e. the angles of the current pose the Nao robot is performing, $\boldsymbol{\theta}_d$

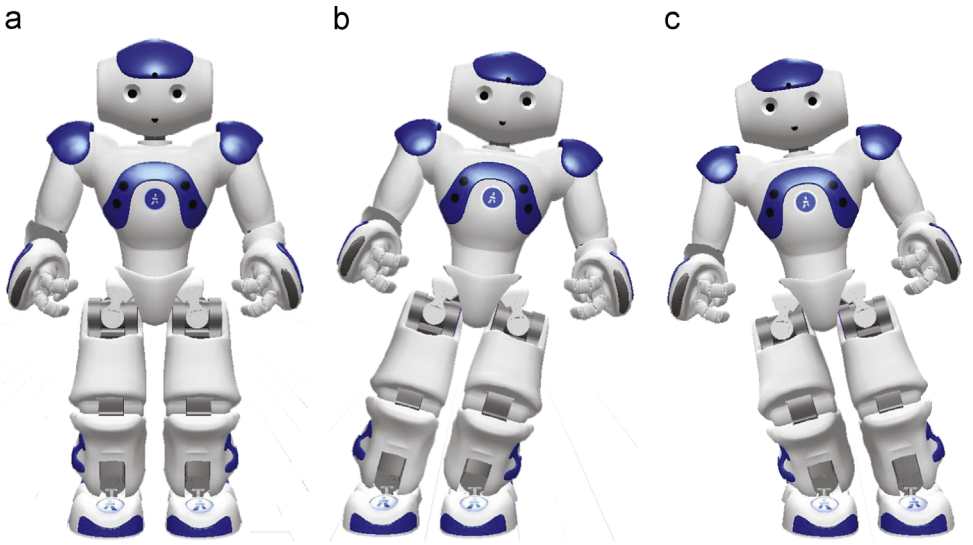

Fig. 3. Transient poses. (a) StandInit. (b) LeftLeaning. (c) RightLeaning.

as the target angle vector, and $\boldsymbol{\theta}_c$ as the corrected angle vector, i.e., the angles obtained after balance control. To be simple, the superscript N for Nao is ignored in this section. Following [22], we can calculate the Jacobian matrix $\boldsymbol{J}_G$ which represents the relations between COM and all the DOF angles. By transferring $\boldsymbol{\theta}_d$ to the Nao robot, a support polygon is obtained. If the projection of $\boldsymbol{X}^d_{COM}$ does not lie within the support polygon, the robot is going to fall. We can make the support leg fixed and let the projection of the COM position be the center of the support polygon, thus getting a corrected COM position $\boldsymbol{X}^c_{COM}$. Now the problem is converted into solving the correction values for all angles, i.e. $\Delta\boldsymbol{\theta}_c$. Then we have

$$\Delta\boldsymbol{X}_{COM} = \boldsymbol{J}_G \Delta\boldsymbol{\theta}_c \tag{3}$$

where $\Delta\boldsymbol{X}_{COM} = \boldsymbol{X}^c_{COM} - \boldsymbol{X}^r_{COM}$ and $\Delta\boldsymbol{\theta}_c = \boldsymbol{\theta}_c - \boldsymbol{\theta}_r$.

Since $\boldsymbol{J}_G$ is not square matrix, the solution to Eq. (3) is not exclusive. As we need the error between $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_d$ as small as possible, the question can be interpreted to a quadratic problem

$$\begin{aligned} &\min \quad \tfrac{1}{2}(\Delta\boldsymbol{\theta}_d - \Delta\boldsymbol{\theta}_c)^T \boldsymbol{W} (\Delta\boldsymbol{\theta}_d - \Delta\boldsymbol{\theta}_c) \\ &\text{s.t.} \quad \boldsymbol{J}_G \Delta\boldsymbol{\theta}_c = \Delta\boldsymbol{X}_{COM} \end{aligned} \tag{4}$$

where $\Delta\boldsymbol{\theta}_d = \boldsymbol{\theta}_d - \boldsymbol{\theta}_r$ and $\boldsymbol{W}$ is a weighting matrix. We can rewrite Eq. (4) as follows:

$$\begin{bmatrix} \boldsymbol{W} & \boldsymbol{J}_G^T \\ \boldsymbol{J}_G & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\theta}_c \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}\Delta\boldsymbol{\theta}_d \\ \Delta\boldsymbol{X}_{COM} \end{bmatrix} \tag{5}$$

where $\boldsymbol{\lambda}$ is the co-state matrix of $\Delta\boldsymbol{\theta}_c$. Solving Eq. (5), we get

$$\begin{aligned} \Delta\boldsymbol{\theta}_c = \Delta\boldsymbol{\theta}_d + \boldsymbol{W}^{-1}\boldsymbol{J}_G^T (\boldsymbol{J}_G \boldsymbol{W}^{-1} \boldsymbol{J}_G^T)^{-1} \\ \times (\boldsymbol{J}_G \Delta\boldsymbol{\theta}_d - \Delta\boldsymbol{X}_{COM}). \end{aligned} \tag{6}$$

Now, we could achieve the stable pose given $\boldsymbol{W}$.

The pose sequence balance control focuses on the fact that there may not exist feasible solution for Eq. (5) when the current pose and the target pose are supported by different legs. This is caused by the physical constraints of the Nao robot, making it unable to find a safe way to perform the transfer. In order to solve the problem, we insert a transient pose between the two poses according to the support leg of the current pose. In this way, the original transfer is spitted into two instead. For instance, suppose we have two contiguous poses in a sequence $\boldsymbol{\theta}_{t1}$ and $\boldsymbol{\theta}_{t2}$ for imitation. However, $\boldsymbol{\theta}_{t1}$ and $\boldsymbol{\theta}_{t2}$ have different support leg situations, assume that they are *Legs* supporting and *LLeg* supporting respectively. According to Section 2.3, the "LeftLeaning" pose should be inserted between the two poses for the reasons mentioned above. Therefore, we replace the desired transfer process with transfer from $\boldsymbol{\theta}_{t1}$ to $\boldsymbol{\theta}_{LeftLeaning}$ at first, and then transfer from $\boldsymbol{\theta}_{LeftLeaning}$ to $\boldsymbol{\theta}_{t2}$.

The whole pose transfer process with balance control is summarized in Algorithm 1.

Algorithm 1. PoseTransferWithBalanceControl.

Input:
- The 3D skeleton data obtained from the Kinect, $Pos = \{Pos_{joint}\}$.

Output:
- Stable Nao robot pose $\boldsymbol{\theta}^N = \{\theta^N_d\}$.
- Perform a transfer from the current pose to $\boldsymbol{\theta}^N$ on the Nao robot.

Get the support leg (*SL*) from *Pos* (Eq. (2)).
Get the human pose $\boldsymbol{\theta}^H$ from *Pos* (Section 2.1).
Get the target Nao pose $\boldsymbol{\theta}^N_{target}$ from $\boldsymbol{\theta}^H$ (Section 2.4).
if $SL == LSL$ (The support leg of the last transferred pose) **then**
  Get the stable pose $\boldsymbol{\theta}^N$ from $\boldsymbol{\theta}^N_{target}$ (Section 2.4).
  Transfer pose $\boldsymbol{\theta}^N$ to the Nao robot.
else
  Load transient pose $\boldsymbol{\theta}^N_{transite}$ based on *LSL* (Section 2.3).
  Transfer pose $\boldsymbol{\theta}^N_{transite}$ to the Nao robot.
  Get the stable pose $\boldsymbol{\theta}^N$ from $\boldsymbol{\theta}^N_{target}$.
  Transfer pose $\boldsymbol{\theta}^N$ to the Nao robot.
end if
$LSL = SL$
return $\boldsymbol{\theta}^N$

3. Pose similarity metric learning

3.1. Motion space

As mentioned before, we believe the motion spaces of humans and Nao robots are different and it is inappropriate to compare human pose with robot pose directly. The reasons can be roughly concluded as follows:

- The bones of humans are pliable while those of Nao robots are not.
- The weight distributions of the two agents are different.
- Compare with Nao robots, humans are better at coordinating the whole body to keep balance, thus getting more flexibility.

Here is the question will be asked, that is "How can we express the motion space for an agent?". With the high dimension of DOFs and complex physical constraints, it is hard to model the motion space explicitly. As a consequence, we choose to model it implicitly.

Similar to sparse coding [23], we choose N human poses as anchor points in the motion space, thus other poses could be a representation of them. Meanwhile, by setting the Nao robot in the animation model, an operator could change its DOF values according to a human pose and record the pose. Then the corresponding (similar) N robot poses are generated.

To make the chosen poses be representative as possible, we design the poses in the full range of an agent (refer to as boundary poses) by considering the static DOF domains. After that, a number of intermediated poses from the initial pose to each boundary pose are selected (Fig. 4).

3.2. Shared latent space

Now, we have a representation of each agent motion space and the correspondence between the N anchor point pairs. Our objection is to give a quantitative evaluation to define the similarity between a human pose and a robot pose.

To begin with, we note that when a human is asked to evaluate the similarity, he/she pays more attention to several DOFs than others in different poses. For example, given a human pose and a Nao robot pose showing standing, many people would think a good correspondence

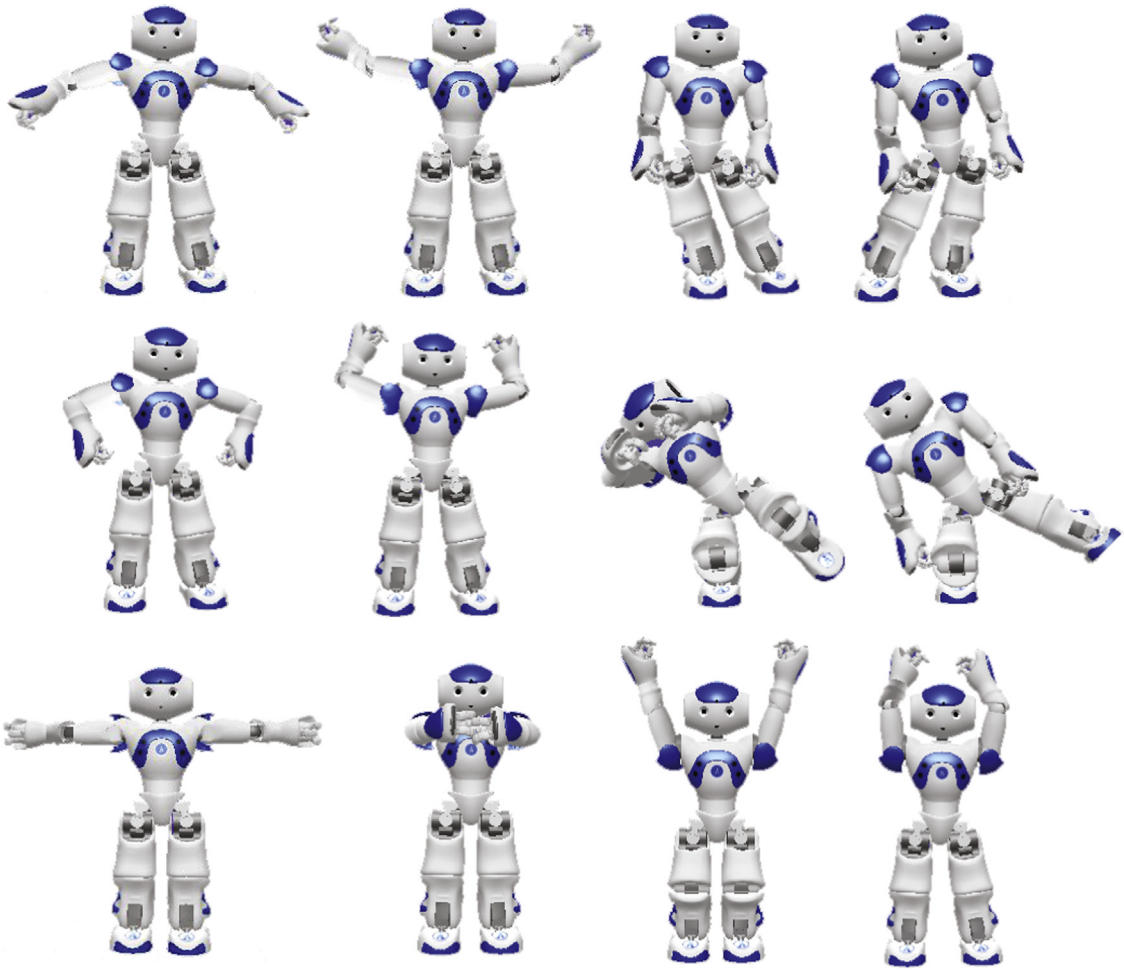

Fig. 4. Some chosen anchor poses of Nao.

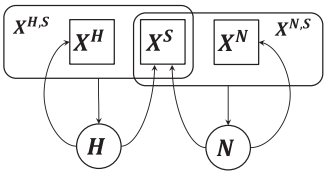

Fig. 5. Shared latent space model.

of "LKneePitch" is more important than "LShoulderRoll". Based on the phenomenon, we have an idea that the human motion space and the Nao robot motion space can be reduced to a combination of a shared space and a personal space. The hyper plane dimensions in the shared space are more discriminative in determining the pose similarity while those in the personal space are less discriminative. Then the problem can be regarded as a dimension reduction problem.

To be further, the original motion spaces can be two observations of the shared space. Inspired by the work in [24,25], the whole shared latent space model is shown in Fig. 5. Denote $\boldsymbol{H}$ as the human motion space, $\boldsymbol{N}$ as the Nao robot motion space and $\boldsymbol{X}^S$ as the shared latent space. $\boldsymbol{X}^H$ and $\boldsymbol{X}^N$ are the private spaces. Thus, the original $\boldsymbol{H}$ and $\boldsymbol{N}$ are reduced to the lower dimension space $\boldsymbol{X}^{H,S}$ and $\boldsymbol{X}^{N,S}$ respectively.

The aim of our model is to find a shared latent representation $\boldsymbol{X}=\{\boldsymbol{X}^H,\boldsymbol{X}^S,\boldsymbol{X}^N\}$ that relates corresponding pairs of $\boldsymbol{H}$ and $\boldsymbol{N}$. Following [26], a shared latent space between the two motion spaces is learned by using the shared GP-LVM (Gaussian process latent variable model), which is modified from the GP-LVM [27] to learn separate sets of Gaussian Processes of different observation spaces. The latent space is learned by maximizing the joint marginal likelihood of the two observation spaces:

$$P(\boldsymbol{H},\boldsymbol{N}|\boldsymbol{X},\Phi_S)=P(\boldsymbol{H}|\boldsymbol{X},\Phi_H)P(\boldsymbol{N}|\boldsymbol{X},\Phi_N) \tag{7}$$

where Φ_H and Φ_N are the hyper-parameters in each GP-LVM and $\Phi_S=\{\Phi_H,\Phi_N\}$. All the important notations used in this paper are listed in Table 2.

Given new poses of each motion space, $\boldsymbol{\theta}^H_{test}$ and $\boldsymbol{\theta}^N_{test}$, we could find their representations in the shared space by the model. The similarity distance of the two poses is then compared in $\boldsymbol{X}^S$ space, normalized by its dimensions

$$Dis\left(\boldsymbol{\theta}^H_{test},\boldsymbol{\theta}^N_{test}\right)=\frac{\|\boldsymbol{X}^S_{H,test}-\boldsymbol{X}^S_{N,test}\|_2}{\|\boldsymbol{X}^S\|}. \tag{8}$$

4. Experimental results

The values of parameters used in the experiment are summarized in Table 3. We record a human pose sequence of 1066 frames using one Kinect at a rate of 0.1 s per frame. The Kinect device can provide color frames, depth frames and skeleton frames in ordinary scenes. The 3D joint positions of the skeleton data are used in our framework as an input, i.e., *Pos*. Considering the movement speed of the Nao robot, we preform the imitation with an interval rather than continuous frames. In the qualitative analysis, a sequence of 97 poses (refer to as *Data* 97) is sampled uniformly from the original sequence, with ignoring the unstable frames at the beginning. In the similarity metric evaluation, another sequence of 450 poses (refer to as *Data* 450) is sampled in the same way with smaller frame interval.

Table 2
Important notations used in this paper.

Notation	Description	Ref.
Pos	Skeleton data (3D positions)	Eq. (1)
SL	Support leg	Section 2.2
LSL	The support leg of the last transferred pose	Algorithm 1
$\boldsymbol{\theta}^H$	Human pose (angles of 20 DOFs)	Section 2.1
$\boldsymbol{\theta}^N$	Nao pose (angles of 24 DOFs)	Section 2.1
$\boldsymbol{\theta}^N_{transite}$	A transient pose	Section 2.3
$\boldsymbol{\theta}^N_r$ or $\boldsymbol{\theta}_r$	The current Nao pose	Section 2.4
$\boldsymbol{\theta}^N_d$ or $\boldsymbol{\theta}_d$	The target Nao pose	Section 2.4
$\boldsymbol{\theta}^N_c$ or $\boldsymbol{\theta}_c$	The corrected Nao pose	Section 2.4
$\boldsymbol{J}_G$	The Jacobian matrix	Section 2.4
$\boldsymbol{X}^r_{COM}$	The COM position of $\boldsymbol{\theta}_r$	Section 2.4
$\boldsymbol{X}^d_{COM}$	The COM position of $\boldsymbol{\theta}_d$	Section 2.4
$\boldsymbol{X}^c_{COM}$	The COM position of $\boldsymbol{\theta}_c$	Section 2.4
$\boldsymbol{H}$	The human motion space	Section 3.2
$\boldsymbol{N}$	The Nao robot motion space	Section 3.2
$\boldsymbol{X}^H$	The private latent space of human motion	Section 3.2
$\boldsymbol{X}^N$	The private latent space of Nao robot motion	Section 3.2
$\boldsymbol{X}^S$	The shared latent space of both agents	Section 3.2

Table 3
Values of parameters in the experiment.

Parameter	Value	Ref.
$\|\boldsymbol{\theta}^H\|$	20	Section 2.1
$\|\boldsymbol{\theta}^N\|$	24	Section 2.1
λ	35 mm	Eq. (2)
$\boldsymbol{W}$	$\boldsymbol{I}$ (Identity)	Eq. (4)
N	50	Section 3.1
$\|\boldsymbol{X}^H\|$	2	Fig. 5
$\|\boldsymbol{X}^N\|$	2	Fig. 5
$\|\boldsymbol{X}^S\|$	10	Fig. 5

4.1. Qualitative analysis

The qualitative experiment is conducted on *Data* 97, some of the imitation results that are varied in support legs and motion ranges are shown in Fig. 6. As we can observe, the balance control is devoted to maintain balance by keeping the actual pose of the Nao robot be similar to the target human pose.

To be further, we make a comparison between the direct imitation poses and the balanced poses to indicate the balance control is effective, as shown in Fig. 7. Meanwhile, we demonstrate the angle trajectories of four DOFs (*RHipPitch*, *RHipRoll*, *RKneePitch*, *RAnklePitch*) in Fig. 8. As observed from Fig. 8, we aim at keeping balance with ensuring the change between the corrected pose and the target pose as small as possible.

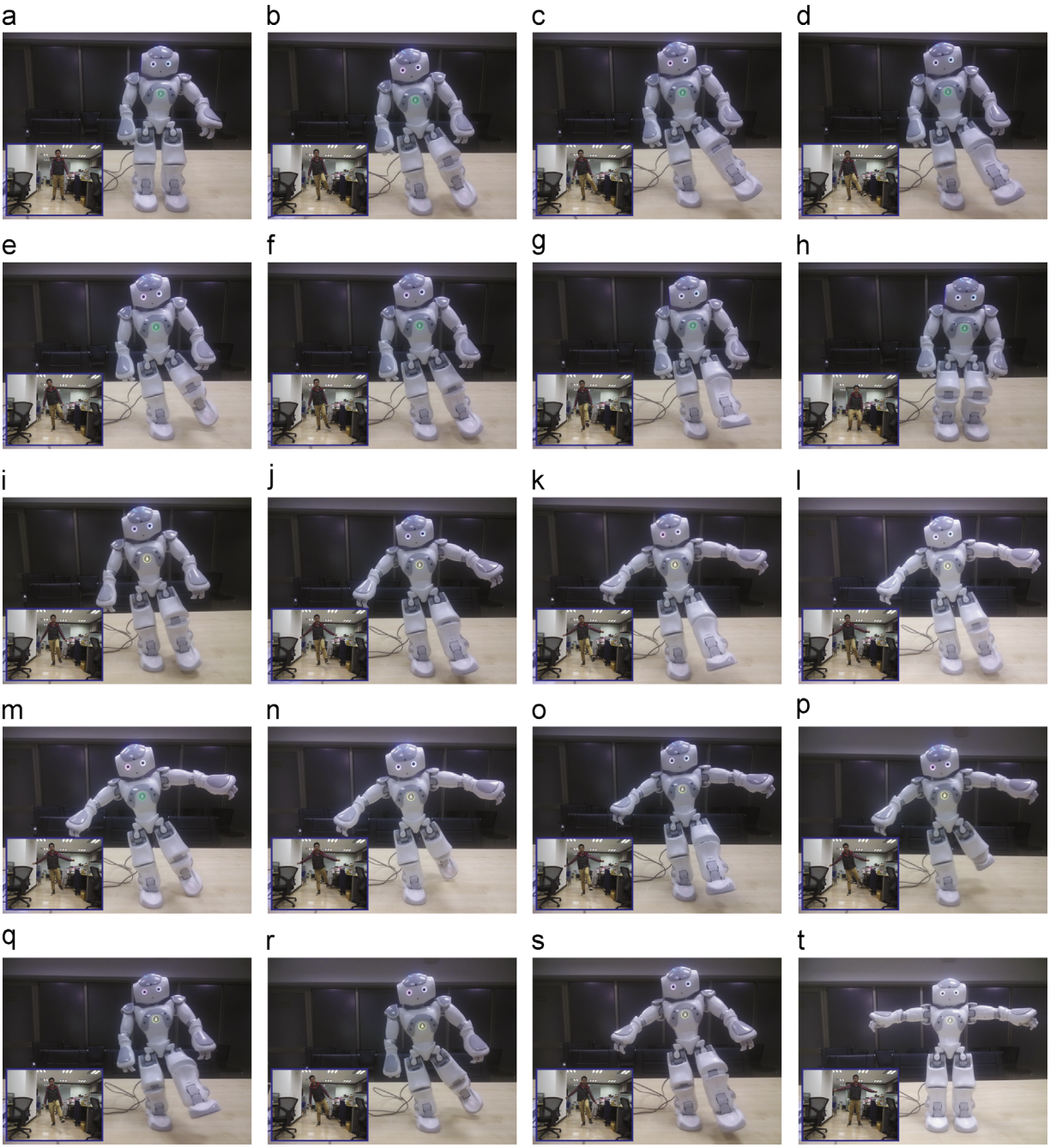

Fig. 6. Some imitation results.

4.2. Quantitative analysis

As mentioned before, asking a person to evaluate the quality of works of imitation is hard to be quantitative. We use the similarity metric described in Section 3 to calculate an accumulated distance for *Data* 97

$$Dis\{Data97\} = \sum_{frame} Dis(\boldsymbol{\theta}^{H}_{frame}, \boldsymbol{\theta}^{N}_{frame}) \tag{9}$$

which equals to 5.58817 in the experiment, then the average single pose similarity distance is 0.05761.

4.3. Similarity metric evaluation

To further investigate the validity of similarity metric, we select a human pose and a few Nao robot poses to be evaluated, both from *Data* 97. The process is repeated for several times with varied test cases. The candidate Nao robot poses are chosen by a human according to their differences from the human poses, including the corresponding Nao pose obtained from our framework. We aim to see if the similarity metric accords with the subjective judgement of humans. An example is shown in Fig. 9. As

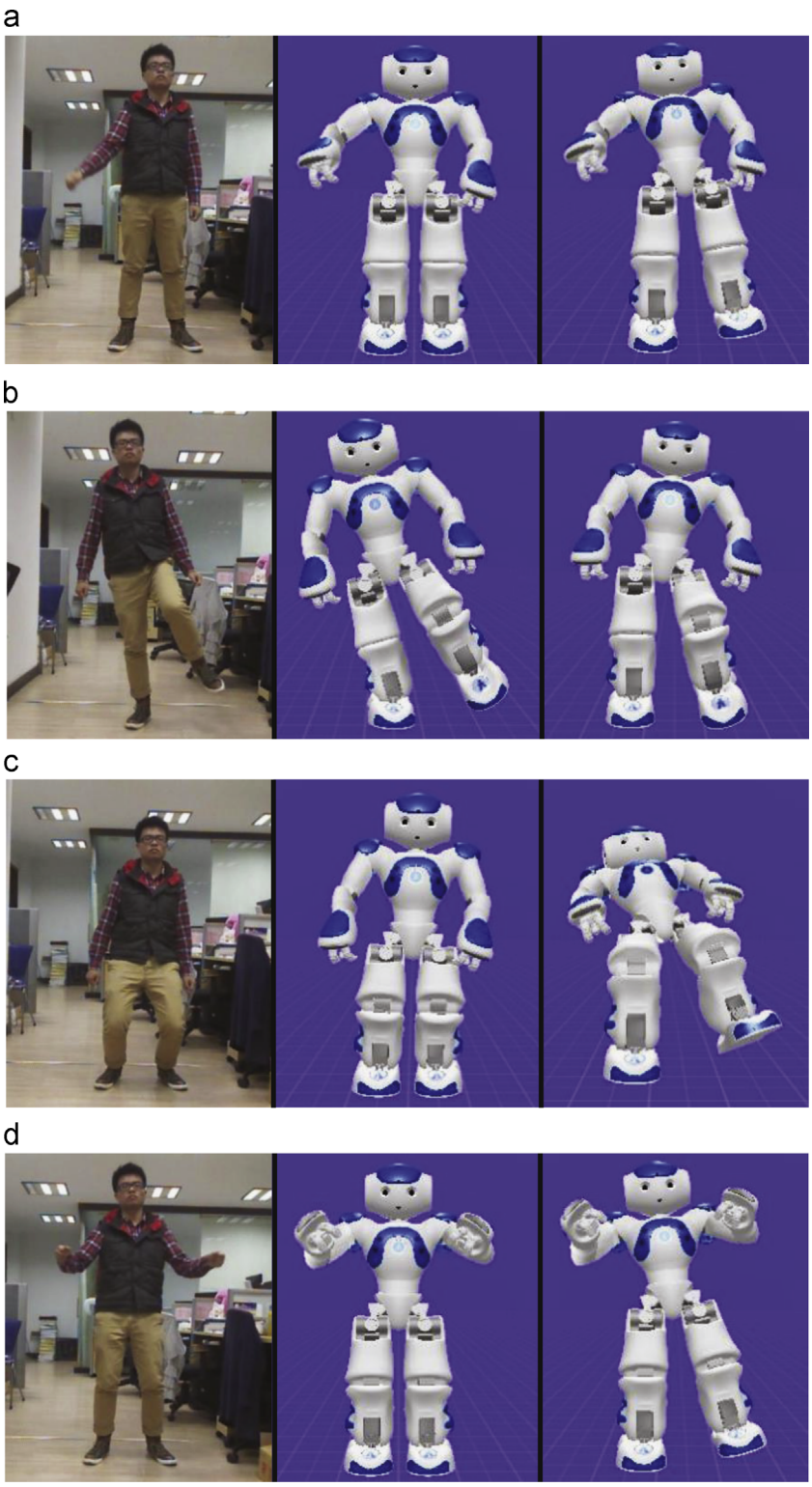

Fig. 7. Comparisons of the balanced poses with the direct imitation poses. The first column shows human poses, while the second and the third columns show the balanced poses and the direct imitation poses respectively.

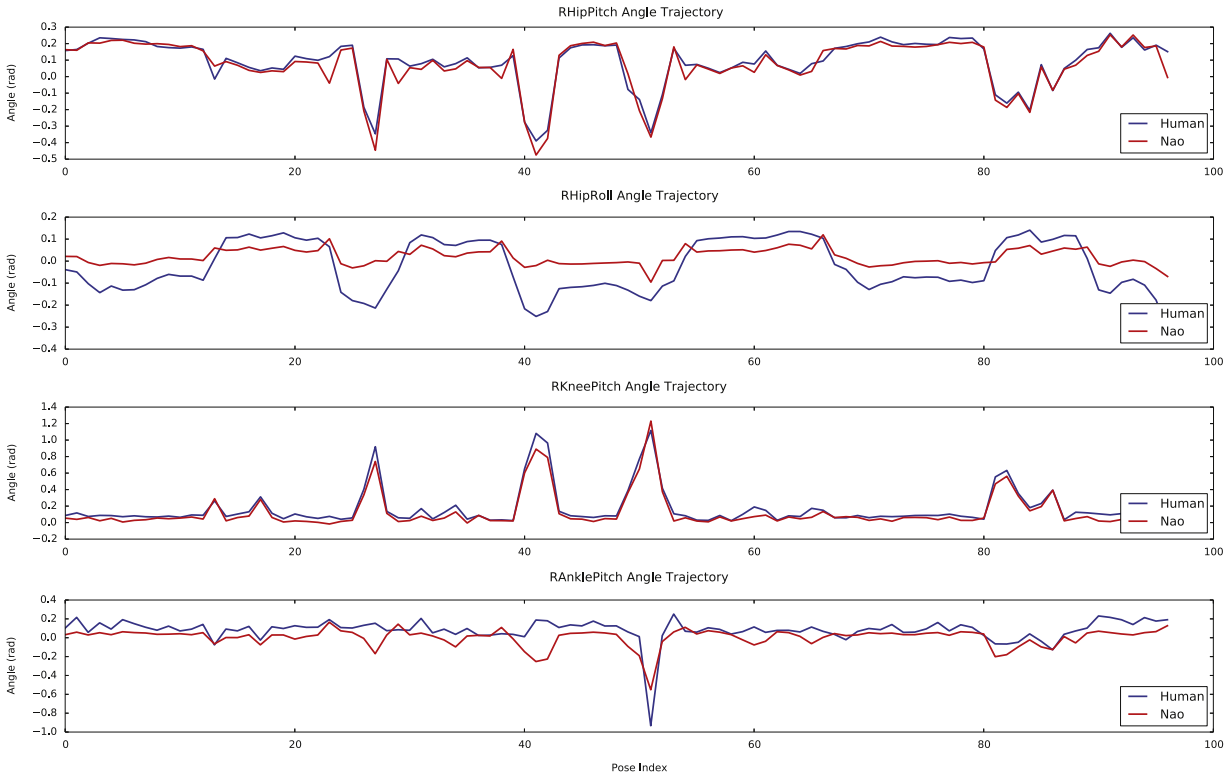

Fig. 8. Angle trajectories of four DOFs (*RHipPitch, RHipRoll, RKneePitch, RAnklePitch*). The blue and red curves show the angles of human and Nao robot respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

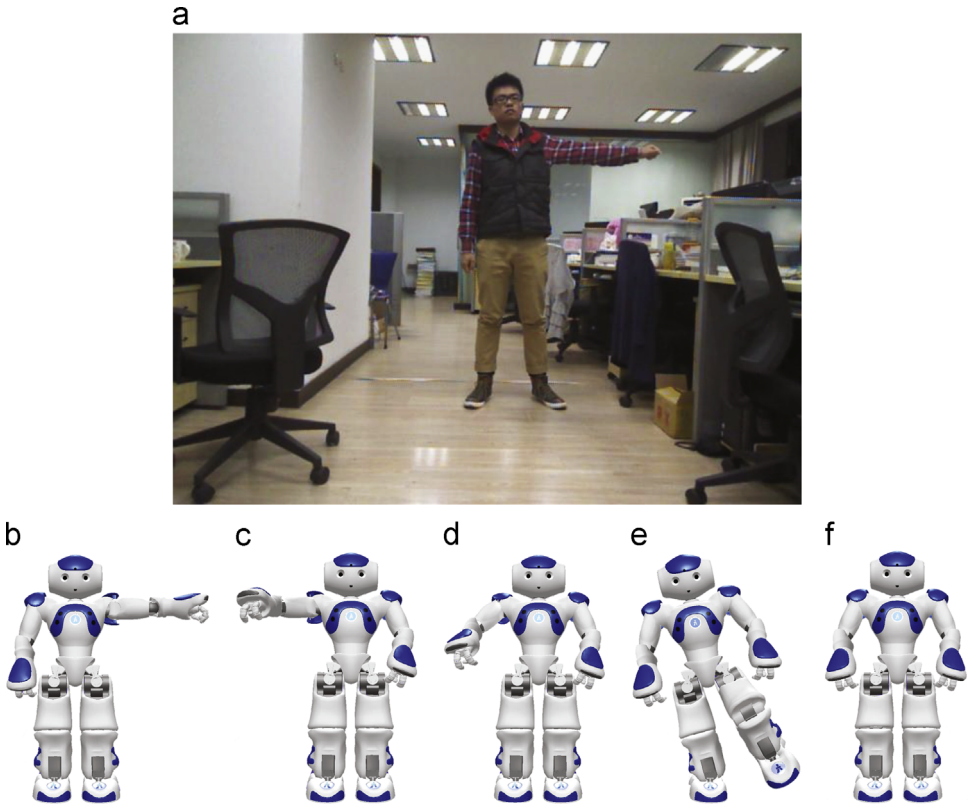

Fig. 9. Compare a human pose with the selected Nao robot poses, where (b) shows the corresponding Nao pose obtained by our imitation framework. The distances between (a) and each (b)–(f) are 0.090922, 0.167751, 0.147825, 0.282243 and 0.133542 respectively, measured in the shared latent space.

Table 4
The discrimination of the similarity metric in local neighbors.

K	*R*	*Data* 97 (%)	*Data* 450 (%)
1	5	61.86	69.11
2	5	71.13	74.67
3	5	85.57	81.33
1	10	57.73	62.00
2	10	63.92	65.78
3	10	71.13	71.33

expected, the corresponding Nao pose (Fig. 9b) achieves the best similarity. Meanwhile, the symmetric Nao robot pose (Fig. 9c) is distinguished from the human pose (Fig. 9a). Moreover, since the poses shown in Fig. 9d and f are similar, their distances to the human pose are approximate and the least similar pose (Fig. 9e) has the furthest distance.

Finally, an experiment is conducted on *Data* 97 and *Data* 450. For each human pose $\boldsymbol{\theta}_i^H$, we refer to the corresponding Nao pose obtained by our framework as $\boldsymbol{\theta}_i^N$. Then we take a range of R Nao robot poses before and after $\boldsymbol{\theta}_i^N$ into consideration in order to know if $\boldsymbol{\theta}_i^N$ is among the nearest K poses of $\boldsymbol{\theta}_i^H$ in the $(2R+1)$ poses in total. To be noted, the frame interval of *Data* 450 is about 0.2 s, thus it maybe unable to discriminate between the contiguous poses for humans. By applying different R and K, the result is summarized in Table 4. It can be concluded that the similarity metric has a stable ability to distinguish the pose from similar neighbors.

5. Conclusion

In this paper, we propose a novel framework for humanoid robot imitation with pose similarity metric learning. DOF angles are used to represent poses. Given a human pose, we adopt the related angles as the target pose of a Nao robot. Through whole body balance control, the stable pose is achieved. To solve the physical constraints of the Nao robot, we apply three transient poses to the original pose transfer, thus making some failure cases feasible. To be further, a latent structure model is applied to study the shared information between human motion space and Nao robot motion space, where the similarity metric is learned. Experimental results demonstrate that the imitation is satisfied and the similarity metric is discriminative.

Regarding to future works, we would like to explore a safe way to deal with self-occluded and auto-collision poses and make use of motion segmentation algorithms to find the key poses to be transferred, thus improving the smooth of the movements.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (61170142), by the National Key Technology R&D Program under Grant 2011BAG05B04, and by the Program of International S&T Cooperation (2013DFG12840).

References

[1] M. Lopes, F. Melo, L. Montesano, J. Santos-Victor, Abstraction levels for robotic imitation: overview and computational approaches, in: From Motor Learning to Interaction Learning in Robots, Springer, 2010, pp. 313–355.
[2] A.R. Ibrahim, W. Adiprawita, Analytical upper body human motion transfer to naohumanoid robot, Int. J. Electr. Eng. Inf. 4 (4) (2012).
[3] M. Riley, A. Ude, K. Wade, C.G. Atkeson, Enabling real-time full-body imitation: a natural way of transferring human movement to humanoids, in: IEEE International Conference on Robotics and Automation, 2003. Proceedings. ICRA'03, vol. 2, IEEE, New York, 2003, pp. 2368–2374.
[4] C. Ott, D. Lee, Y. Nakamura, Motion capture based human motion recognition and imitation by direct marker control, in: Eighth IEEE-RAS International Conference on Humanoid Robots, 2008. Humanoids 2008, IEEE, New York, 2008, pp. 399–405.
[5] J. Aleotti, A. Skoglund, T. Duckett, Position teaching of a robot arm by demonstration with a wearable input device, in: International Conference on Intelligent Manipulation and Grasping (IMG04), 2004, pp. 1–2.
[6] C. Stanton, A. Bogdanovych, E. Ratanasena, Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning, in: Proceedings of the Australasian Conference on Robotics and Automation, 2012.
[7] S. Wehner, M. Bennewitz, Humanoid gait optimization based on human data, Automat.: J. Control Meas. Electron. Comput. Commun. 52 (3) (2011).
[8] T. Sugihara, Y. Nakamura, H. Inoue, Real-time humanoid motion generation through zmp manipulation based on inverted pendulum control, in: IEEE International Conference on Robotics and Automation, 2002. Proceedings. ICRA'02, vol. 2, IEEE, New York, 2002, pp. 1404–1409.
[9] B. Stephens, C. Atkeson, Modeling and control of periodic humanoid balance using the linear biped model, in: Ninth IEEE-RAS International Conference on Humanoid Robots, 2009. Humanoids 2009, IEEE, New York, 2009, pp. 379–384.
[10] J. Koenemann, M. Bennewitz, Whole-body imitation of human motions with a nao humanoid, in: 2012 Seventh ACM/IEEE International Conference on Human–Robot Interaction (HRI), IEEE, New York, 2012, pp. 425–425.
[11] F. Zuher, R. Romero, Recognition of human motions for imitation and control of a humanoid robot, in: Robotics Symposium and Latin American Robotics Symposium (SBR-LARS), 2012 Brazilian, IEEE, New York, 2012, pp. 190–195.
[12] T. Harada, S. Taoka, T. Mori, T. Sato, Quantitative evaluation method for pose and motion similarity based on human perception, in: 2004 Fourth IEEE/RAS International Conference on Humanoid Robots, vol. 1, IEEE, New York, 2004, pp. 494–512.
[13] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, J. Xiao, Learning a 3d human pose distance metric from geometric pose descriptor, IEEE Trans. Vis. Comput. Graph. 17 (11) (2011) 1676–1689.
[14] J. Yu, Y. Rui, Y.Y.Tang, D. Tao, High-order distance-based multiview stochastic learning in image classification, IEEE Transactions on Cybernetics, http://dx.doi.org/10.1109/TCYB.2014.2307862, 2014.
[15] J. Yu, D. Tao, M. Wang, Adaptive hypergraph learning and its application in image classification, IEEE Trans. Image Process. 21 (7) (2012) 3262–3272.
[16] J. Yu, M. Wang, D. Tao, Semisupervised multiview distance metric learning for cartoon synthesis, IEEE Trans. Image Process. 21 (11) (2012) 4636–4648.
[17] J. Yu, D. Tao, J. Li, J. Cheng, Semantic preserving distance metric learning and applications, Inf. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.025.
[18] P. Li, M. Wang, J. Cheng, C. Xu, H. Lu, Spectral hashing with semantically consistent graph for image indexing, IEEE Trans. Multimed. 15 (1) (2013) 141–152.
[19] M. Wang, R. Hong, X.-T. Yuan, S. Yan, T.-S. Chua, Movie2comics: towards a lively video content presentation, IEEE Trans. Multimed. 14 (3) (2012) 858–870.
[20] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, Assistive tagging: a survey of multimedia tagging with human–computer joint exploration, ACM Comput. Surv. (CSUR) 44 (4) (2012) 25.
[21] N. Naksuk, C.G. Lee, S. Rietdyk, Whole-body human-to-humanoid motion transfer, in: 2005 Fifth IEEE-RAS International Conference on Humanoid Robots, IEEE, New York, 2005, pp. 104–109.
[22] T. Sugihara, Y. Nakamura, Whole-body cooperative balancing of humanoid robot using cog Jacobian, in: IEEE/RSJ International

Conference on Intelligent Robots and Systems, 2002, vol. 3, IEEE, New York, 2002, pp. 2575–2580.
[23] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, Adv. Neural Inf. Process. Syst. 19 (2007) 801.
[24] A.P. Shon, K. Grochow, R.P. Rao, Robotic imitation from human motion capture using Gaussian processes, in: 2005 Fifth IEEE-RAS International Conference on Humanoid Robots, IEEE, New York, 2005, pp. 129–134.
[25] V.A. Prisacariu, I. Reid, Shared shape spaces, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, New York, 2011, pp. 2587–2594.
[26] A. Shon, K. Grochow, A. Hertzmann, R.P. Rao, Learning shared latent structure for image synthesis and robotic imitation, in: Advances in Neural Information Processing Systems, 2005, pp. 1233–1240.
[27] C.H. Ek, P.H. Torr, N.D. Lawrence, Gaussian process latent variable models for human pose estimation, in: Machine Learning for Multimodal Interaction, Springer, 2008, pp. 132–143.