Humanoid Robot Imitation with Pose Similarity Metric Learning

Jie Lei*, Mingli Song*, Ze-Nian Li†, Chun Chen*, Xianghua Xu‡ and Shiliang Pu§

*College of Computer Science
Zhejiang University, Hangzhou, China 310027
Email: {ljaylei, brooksong, chenc}@zju.edu.cn
†School of Computing Science
Simon Fraser University, Burnaby, BC, Canada V5A1S6
Email: li@sfu.ca
‡School of Computer Science
Hangzhou Dianzi University, Hangzhou, China 310018
Email: xhxu@hdu.edu.cn
§Hangzhou Hikvision Digital Technology Co., Ltd
Hangzhou, China 310052
Email: pushiliang@hikvision.com

Abstract—Imitation is considered to be a kind of social learning that allows the transfer of information, actions, behaviours, 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.

Index Terms—humanoid robot; imitation; pose transfer; similarity metric

I. 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, etc, 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. In [2], an analytical method was proposed to transfer the upper body motion from human to humanoid robot. Aleotti et al. [3] 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. [4] extended it to a humanoid robot by training a feed forward neural network with particle swarm optimisation for each degree of freedom (DOF). In the data collecting process, a robot was used to lead a human operator through series of paired synchronised movements captured by a motion capture, which was time-consuming and tedious. As they were mentioned, the position of the robot’s ankles did not employ neural networks to ensure robot stability. 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 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 ([5], [6], [7]). In conclusion, these works have the following limitations:

- Imitation of the upper body or a single part is insufficient to meet the needs of humanoid robot ([2], [3]);
- With requiring motion capture equipment, it is expensive for general use and unnatural for human-robot interaction ([3], [4], [8]);
- Lack of balance control and the whole body control ([4]);
- The imitation results are not qualitatively evaluated ([4], [8], [9]).

After performing the pose imitation, another important issue is “how can we evaluate the imitation similarity between a robot slave and the master”. In [9], 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 $L2$, which does not sufficiently utilize the data dependency between DOFs. In [10], different weights were learned for DOFs, in correspondence with the fact that some DOFs had more influences on determining the similarity. Chen et al. [11] 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 constraints, i.e., they have different motion spaces. It is inappropriate to compare them directly.

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 adopts dynamic balance control with real-time 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.

II. HUMANOID ROBOT IMITATION

A. 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 I, which are angles between pairs of related vectors. For example,

$$\theta_{HeadPitch}^H = (DV(\text{PosSpine}, \text{PosShoulderCenter}),\ DV(\text{PosShoulderCenter}, \text{PosHead}))$$

(1)

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 $\theta^H = \{\theta^H_d\}$ where $d$ is the name of a DOF.

The Nao robot owns 26 DOFs (Fig. 1), where we exclude the “LHand” and “RHand” DOFs and choose the rest 24 DOFs, as listed in Table I, to represent a Nao pose, denoted as $\theta^N = \{\theta^N_d\}$. 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.

B. 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 in short).

The support leg should be figured out before controlling the balance. There are three situations, i.e., $L\text{Leg}$, $R\text{Leg}$, $\text{Legs}$, 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

$$SL = \begin{cases} L\text{Leg} : |\text{Pos}_{FootLeft}^Y - \text{Pos}_{FootRight}^Y| \leq \lambda \\
L\text{Leg} : |\text{Pos}_{FootLeft}^Y - \text{Pos}_{FootRight}^Y| > \lambda \\
R\text{Leg} : |\text{Pos}_{FootLeft}^Y - \text{Pos}_{FootRight}^Y| < \lambda \end{cases}$$

(2)

where $SL$ is short for support leg and $\lambda$ is a threshold for smoothing.
C. 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. 2. 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.

D. Balance Control

The 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 $\theta^H$, we choose the 20 corresponding angles and set the rest 4 angles as 0 to form a target pose of the Nao robot. Denote $X_{\text{COM}}$ as the COM position of the robot, $\theta_e$ as the current angle vector, $\theta_d$ as the target angle vector, and $\theta_r$ as the actual angle vector, i.e., the angles obtained after balance control. To be simple, the superscript $N$ for Nao is ignored in this section. Following [12], we can calculate the Jacobian matrix $J_G$ which represents the relations between $X_{\text{COM}}$ and all the DOF angles. By transferring $\theta_d$ to the Nao robot, a support polygon is obtained. If the projection of $X_{\text{COM}}$ to the Nao robot 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 $X_{\text{COM}}^c$. Now the problem is converted into solving the correction values for all angles, i.e. $\Delta \theta_e$. Then we have:

$$\Delta X_{\text{COM}} = J_G \Delta \theta_e$$

where $\Delta X_{\text{COM}} = X_{\text{COM}}^c - X_{\text{COM}}$ and $\Delta \theta_e = \theta_e - \theta_r$.

Since $J_G$ is not square matrix, the solution to Eqn. (3) is not exclusive. As we need the error between $\theta_e$ and $\theta_r$ as small as possible, the question can be interpreted to a quadratic problem

$$\min \frac{1}{2}(\Delta \theta_d - \Delta \theta_e)^T W (\Delta \theta_d - \Delta \theta_e)$$

s.t. $J_G \Delta \theta_e = \Delta X_{\text{COM}}$

where $\Delta \theta_d = \theta_d - \theta_e$ and $W$ is a weighting matrix. We can rewrite Eqn. (4) as follows,

$$\begin{bmatrix} W & J_G^T \\ J_G & 0 \end{bmatrix} \begin{bmatrix} \Delta \theta_e \\ \lambda \end{bmatrix} = \begin{bmatrix} W \Delta \theta_d \\ \Delta X_{\text{COM}} \end{bmatrix}$$

where $\lambda$ is the co-state matrix of $\Delta \theta_e$. Solving Eqn. (5), we get

$$\Delta \theta_e = \Delta \theta_d + W^{-1} J_G^T (J_G W^{-1} J_G^T)^{-1} (J_G \Delta \theta_d - \Delta X_{\text{COM}}).$$

Now, we could achieve the stable pose given $W$.

The pose sequence balance control focuses on the fact that there may not exist feasible solution for Eqn. (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 split into two instead. The whole pose transfer process with balance control is summarized in Algorithm 1.

<table>
<thead>
<tr>
<th>Algorithm 1 PoseTransferWithBalanceControl</th>
</tr>
</thead>
</table>

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

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

1. Get the support leg ($SL$) from $P_{\text{Pos}}$, (Eqn. (2))
2. Get the human pose $\theta^H$ from $P_{\text{Pos}}$, (Sec. II-A)
3. Get the target Nao pose $\theta^N_{\text{target}}$ from $\theta^H$. (Sec. II-D)
4. If $SL == LSL$ (The support leg of the last transfer) then
5. Get the stable pose $\theta^N$ from $\theta^N_{\text{target}}$. (Sec. II-D)
6. Transfer pose $\theta^N$ to the Nao robot.
7. Else
8. Load transient pose $\theta^N_{\text{transite}}$ based on $LSL$. (Sec. II-C)
9. Transfer pose $\theta^N_{\text{transite}}$ to the Nao robot.
10. Get the stable pose $\theta^N$ from $\theta^N_{\text{target}}$.
11. Transfer pose $\theta^N$ to the Nao robot.
12. End if
13. $LSL = SL$
14. Return $\theta^N$

III. POSE SIMILARITY METRIC LEARNING

A. 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 bones of humans are pliable while those of Nao robots are not;
- The bones of humans are pliable while those of Nao robots are not;
- The bones of humans are pliable while those of Nao robots are not;
Fig. 3. Shared latent space model.

- The weight distributions of the two agents are different;
- Compare to 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, 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 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.

B. 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 of “L.KneePitch” is more important than “L.ShoulderRoll”. 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 [13] [14], the whole shared latent space model is showed in Fig. 3. Denote $H$ as the human motion space, $N$ as the Nao robot motion space and $S$ as the shared latent space. $X^H$ and $X^N$ are the private spaces. Thus, the original $H$ and $N$ are reduced to the lower dimension space $X^{HS}$ and $X^{NS}$ respectively.

Following [14], 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 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,

$$P(H, N | X, \Phi_S) = P(H | X, \Phi_H)P(N | X, \Phi_N)$$ (7)

where $\Phi_H$ and $\Phi_N$ are the hyper-parameters in each GP-LVM and $\Phi_S = \{\Phi_H, \Phi_N\}$.

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

$$D_{\text{dist}}(\theta^H_{test}, \theta^N_{test}) = \frac{||X^S_{H, test} - X^S_{N, test}||_2}{||X^S||}.$$ (8)

IV. EXPERIMENTAL RESULTS

The values of parameters used in the experiment are summarized in Table II. We record a human pose sequence of 1066 frames using the Kinect at a rate of 0.1s per frame. In the qualitative analysis, a sequence of 97 poses (refer to as Data97) 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 Data450) is sampled in the same way with smaller frame interval.

A. Qualitative Analysis

The qualitative experiment is conducted on Data97, some of the imitation results that are varied in support legs and motion ranges are showed in Fig. 4. As we can observe, the balance control is devoted to maintain balance with keeping the actual pose of the Nao robot be similar to the target human pose.

B. Quantitative Analysis

As mentioned before, asking a person to evaluate the quality of works of imitation is hard to be quantitative. We use
TABLE III
THE DISCRIMINATION OF THE SIMILARITY METRIC IN LOCAL NEIGHBORS.

<table>
<thead>
<tr>
<th>(R)</th>
<th>(R)</th>
<th>Data97(%)</th>
<th>Data450(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>61.36</td>
<td>69.11</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>71.13</td>
<td>74.67</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>85.57</td>
<td>81.33</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>57.73</td>
<td>62.00</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>63.92</td>
<td>65.78</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>71.13</td>
<td>71.33</td>
</tr>
</tbody>
</table>

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

C. 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 Data97. 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 showed in Fig. 5. As expected, the corresponding Nao pose (Fig. 5(b)) achieves the best similarity. Meanwhile, the symmetric Nao robot pose (Fig. 5(c)) is distinguished from the human pose (Fig. 5(a)). Moreover, since the poses showed in Fig. 5(d) and Fig. 5(f) are similar, their distances to the human pose are approximate and the least similar pose (Fig. 5(e)) has the furthest distance.

Finally, an experiment is conducted on Data97 and Data450. For each human pose \(\theta^H_i\), we refer to the corresponding Nao pose obtained by our framework as \(\theta^N_i\). Then we take a range of \(R\) Nao robot poses before and after \(\theta^N_i\) into consideration in order to know if \(\theta^N_i\) is among the nearest \(K\) poses of the \((2R + 1)\) poses in total. To be noted, the time interval of Data450 is about 0.2s, thus it maybe unable to discriminate between the contiguous poses for humans. By applying different \(R\) and \(K\), the result is summarized in Table III. It can be concluded that the similarity metric has a stable ability to distinguish the pose from similar neighbors.

V. 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.
Fig. 5. 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) to (f) are 0.090922, 0.167751, 0.147825, 0.282243 and 0.133542 respectively, measured in the shared latent space.

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 movements.

ACKNOWLEDGMENT

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 (2013D-FG12840).

REFERENCES


