Prepare the Chair for the Bear! Robot Imagination of Sitting Affordance to Reorient Previously Unseen Chairs Supplementary Material

Xin Meng[1], Hongtao Wu[2], Sipu Ruan[1], Gregory S. Chirikjian[1]

A. Mathmatical Details

1) Rotation Enumeration: In Sec. IV-A, we define that two rotations $R_1 \in SO(3)$ and $R_2 \in SO(3)$ are functionally and perceptibly equivalent if there exists an $\alpha \in [0, 2\pi)$ such that $R_2 = R_z(\alpha)R_1$. When combined with a position $\mathbf{p} = [x, y, z]^T \in \mathbb{R}^3$, they form two poses, $g_1 = (R_1, \mathbf{p})$ and $g_2 = (R_2, \mathbf{p})$, that are functionally and perceptibly equivalent, *i.e.*, $g_2 = hg_1$, where $h = (R_z(\alpha), [0, 0, 0]^T) \in H$. We claim that for any pose $g = (R, \mathbf{p}) \in SE(3)$, there exists a rotation $R' = R_{ZYX}(0, \beta', \gamma')$ such that $g' = (R', \mathbf{p})$ is functionally and perceptibly equivalent to g, where $\beta' \in [0, 2\pi)$, $\gamma' \in [0, 2\pi)$. We prove this by proving that a functionally and perceptibly equivalent rotation $R_{ZYX}(0, \beta', \gamma')$ exists for any $R \in SO(3)$. R can be decomposed using Euler angles as $R = R_z(\alpha)R_y(\beta)R_x(\gamma)$. The rotation $R' = R_y(\beta)R_x(\gamma)$ is functionally and perceptibly equivalent to R given that $R' = R_z(-\alpha)R$. Thus, by enumerating β and γ, we are able to cover all cosets in $SO(3)$ and enumerate a rotation that is functionally and perceptibly equivalent to any R.

For any stable pose $g_s = (R_s, \mathbf{p}_s)$, if we set the object rotation as R_s and drop it from above the plane, the object will land in a pose g that is close to and functionally equivalent to g_s. In the rotation enumeration, we vary β and γ from $[0, 2\pi)$ in discrete increments $\Delta\beta$ and $\Delta\gamma$. This enumeration is able to find a rotation that is functionally and perceptibly equivalent to (or close enough to a rotation that is functionally and perceptibly equivalent to) any rotation, including the rotation R_s of any stable pose g_s. By dropping the object in this rotation, the object will be very likely to result in a pose that is functionally and perceptibly equivalent to g_s. That is, by dropping the object from all the enumerated rotations, we are able to find a set of all functionally and perceptibly unique stable poses G_s. The set of cosets formed by all the found stable poses $S = \{Hg_s | g_s \in G_s\}$ contains all the stable pose cosets of the object.

2) Perceptible Difference: Given two random poses $g_1 = (R_1, \mathbf{p}_1) \in SE(3)$ and $g_2 = (R_2, \mathbf{p}_2) \in SE(3)$, we claim that the relative rotation matrix $R_{12} = R_2 R_1^T$ can be decomposed as $R_{12} = R_z(\alpha_{12})R_{xy}(\phi_{12})$ where $R_{xy}(\cdot)$ is a rotation about an axis in the xy-plane and ϕ_{12} is the rotation angle. Moreover, we claim that the perceptible difference $d_{pcp} = \phi_{12}$.

Proof:

A rotation can also be expressed in exponential coordinates:

$$R = \exp(\widehat{\mathbf{w}}\theta) \tag{1}$$

$\mathbf{w} \in \mathbb{R}^3$ ($\widehat{\mathbf{w}} \in so(3)$) is a unit vector indicating the axis of the rotation; θ is the rotation angle.

A rotation R can be parameterized with ZYZ Euler angles:

$$\begin{aligned} R &= R_Z(\alpha_1)R_Y(\beta_2)R_Z(\alpha_3) \\ &= R_Z(\alpha_1+\alpha_3)R_Z(-\alpha_3)R_Y(\beta_2)R_Z(\alpha_3) \\ &= R_Z(\alpha_1+\alpha_3)R_Z(-\alpha_3)\exp(\widehat{\mathbf{n}}_y\beta_2)R_Z^T(-\alpha_3) \\ &= R_Z(\alpha_1+\alpha_3)\exp\left[R_Z(-\alpha_3)\widehat{\mathbf{n}}_y\beta_2 R_Z^T(-\alpha_3)\right] \\ &= R_Z(\alpha_1+\alpha_3)\exp\left[(\widehat{R_Z(-\alpha_3)\mathbf{n}_y})\beta_2\right] \end{aligned} \tag{2}$$

$\exp\left[(\widehat{R_Z(-\alpha_3)\mathbf{n}_y})\beta_2\right]$ is the exponential representation of a rotation. β_2 is the rotation angle. $R_Z(-\alpha_3)\mathbf{n}_y$ indicates the rotation axis in xy-plane in which $\mathbf{n}_y = [0, 1, 0]^T$ is the y-axis unit vector and $R_Z(-\alpha_3)$ is a rotation about the z-axis of the world frame. Therefore, R_{12} can be parameterized by a rotation about the z-axis and a rotation about an axis in the xy-plane, of which the rotation angle β_2 can be computed from $\arccos(r_{33})$ from the matrix representation of R_{12}.

To facilitate understanding, let's consider a virtual frame attached to the object, Fig. 1(b) shows an example. Its rotation is identity $R = I_3$ which is the same as that of the world frame. When the object is rotated by R_{12} in the world frame, the rotation of this frame becomes R_{12}. The last column of R_{12} indicates the unit vector of the z-axis of the virtual frame as seen in the world frame. r_{33} is the dot product of the unit vector of the virtual frame z-axis and the unit vector of the world frame z-axis $\mathbf{n}_z = [0, 0, 1]^T$. That is, r_{33} is the cosine value of the angle between these two z-axes and d_{pcp} is the value of the angle. The concept of *perceptible difference* can be understood as the change in the orientation of the z-axis of this virtual frame after a rotation (Fig. 1(c)). It reflects the change of an object's perceptible (or unoccluded) section when it is reoriented. The section does not change regardless of the rotation about the z-axis of the world frame. That is, the perceptibility of the object does not change if $r_{33} = 0$. By examining $d_{pcp} = \arccos(r_{33})$, we can quantify the change in the perceptibility of the object. The largest perceptible difference between two object poses indicates the most different view that can be captured by the robot.

[1]X. Meng, S. Ruan, and G. S. Chirikjian are with the Department of Mechanical Engineering, National University of Singapore, Singapore. {mengxin, ruansp, mpegre}@nus.edu.sg.

[2]H. Wu is with the Laboratory for Computational Sensing and Robotics (LCSR), Johns Hopkins University, Baltimore, MD 21218, USA hwu67@jhu.edu.

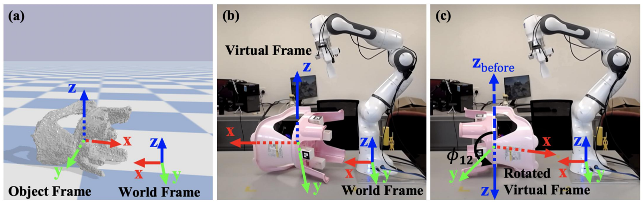

Fig. 1. (a) shows the reconstructed object model and object frame. (b) shows the virtual frame for evaluating perceptible difference in the real-world setting. (c) shows the rotated virtual frame after manipulation, z_{before} shows the z-axis of the virtual frame before manipulation, the value of the angle between these two z-axes ϕ_{12} is the perceptible difference.

Finally, we claim that for any $h_1 g_1 \in H g_1$ and $h_2 g_2 \in H g_2$, the perceptible difference are the same, *i.e.*, $d_{pcp}(h_1 g_1, h_2 g_2) = d_{pcp}(g_1, g_2)$.

Proof:

$$
\begin{aligned}
d_{pcp}(h_1 g_1, h_2 g_2) &= \arccos[n_z^T (R_z(\alpha_2) R_2 (R_z(\alpha_1) R_1)^T) n_z] \\
&= \arccos[n_z^T R_z(\alpha_2)(R_2 R_1^T) R_z(\alpha_1)^T n_z] \\
&= \arccos[n_z^T (R_2 R_1^T) n_z] \\
&= d_{pcp}(g_1, g_2)
\end{aligned}
$$

where $R_z(\alpha) n_z = n_z$ is used.

B. Experiment Details

1) Imagination: Following the settings of our previous papers, in stable pose imagination, the initial orientation of the object is enumerated by varying β and γ in 15 discrete values within $[0, 2\pi)$, *i.e.*, $\Delta\beta = \Delta\gamma = 2\pi/15$. In each sitting imagination, the agent is dropped onto the chair in 8 different orientations, *i.e.*, $\Delta\alpha = \pi/4$.

The SAM classifies and evaluates the configuration of an agent based on five criteria, which range from coarse to fine. The joint score and link score provide an overall comparison with the key configuration, while the other three focus on different specific aspects. The evaluation of contact points ensures proper support for the agent's body. The symmetry score takes into account the intuition that natural sitting configurations for humans are typically symmetrical. The height score focuses on the seat's height and flatness, ensuring that it is comfortable for human use. The thresholds are: $J_{max} = 2.0$, $L_{max} = 0.5$, $C_{min} = 4$, $S_{max} = 2.0$, $SS_{min} =$, $H_{tmin} = H_{pmin} = 0.15 H_{agent}$, $H_{tmax} = H_{pmax} = 0.15 H_{agent} + 1.0$, $\Delta H_{min} = 0.15 H_{agent}$.

In functional pose prediction, the thresholds $N_{min} = 3$.

In Eq 3, $\Delta v_{es} = 0.99, \Delta z_{es} = 0.01$m.

2) Object Frame: During each object reconstruction, the robot fuses the depth image captured from different viewpoints and reconstructs the object 3D mesh. To establish a frame for the object, we compute the Object Bounding Box (OBB) of the mesh. The object frame is positioned at the geometry center. The z-axis is set parallel to the z-axis of the world frame. The x-axis is parallel to the projection of an OBB edge in the xy-plane. The y-axis is set according to the right-hand rule. Note that the object pose is the pose of the object frame with respect to the world frame. Fig. 1 (a) and (b) shows the object frame in object model and real world.

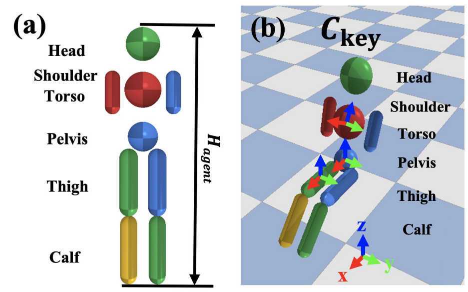

Fig. 2. Agent configuration. (a) shows the link names. (b) shows the key configuration C_{key}.

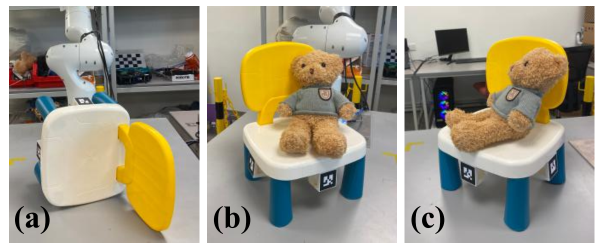

Fig. 3. Failure Case. (a) Initial pose of a test chair. (b) and (c) shows the result of chair preparation and bear seating.

3) Agent Configuration: The human agent is modeled with an articulated human body with the forearms and feet trimmed off. Each link of the body has a frame attached to its center as shown in Fig. 2 (b). The key configuration refers to a pre-defined ideal sitting configuration for the agent, where the agent orientation and the joint angles are set to ideal values. The values are determined based on the definition of *sit*.

4) Results: In the real experiments, our method achieved a 100 % success rate in chair classification and functional pose prediction of the chairs. In the 30 trials of chairs, the robot successfully seats the bear on the chair in 29 trials. The failure case is shown in Fig. 3. Among the imagined sitting poses, the highest-ranked sitting direction is not perpendicular to the back of the chair. The sitting position is difficult for the NAO robot to reach from such a direction. The control error and the restricted workspace of NAO also make it difficult to precisely put the teddy bear into the imagined pose. Thus, the bear only has its head

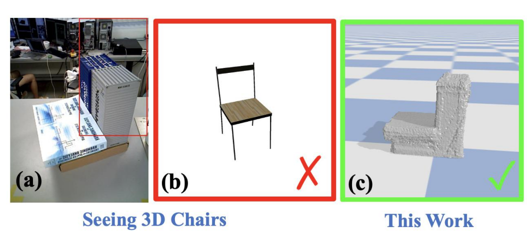

Fig. 4. (a)-(b) A failure example of pose estimation of an upright chair using [1]. (c) Functional pose estimated by our approach.

TABLE I

AVERAGE TIME CONSUMPTION IN REAL ROBOT EXPERIMENTS

	Execution	Time Consumption (s)
Object Complete Reconstruction	1st scan	51
	Reconstruction of incomplete model	33
	Stable pose imagination	13
	Robot manipulation	135
	2nd scan	56
	Reconstruction of complete model	70
Prepare the Chair for the Bear	Stable pose imagination	10
	Sitting imagination	65
	Robot manipulation to reorient the chair	204
Seating the Bear on the Chair		166

supported by the chair back from the side. Three out of five annotators do not regard it as a successful sitting.

We also showcase a failure example of an image-based method [1] for chair detection and pose estimation in Fig. 4, where the object is an improvised chair assembled with books and boxes. This object does not appear to be like a chair but is able to afford the sitting functionality. The method in [1] is not able to detect the object correctly or find the pose of the object. In contrast, our approach classifies it as a chair and is capable of predicting the functional pose of the object.

The imagination and reconstruction were performed on a computer running Intel Core i7-11700K @ 3.6GHz CPU. The Franka robot control modules and NAO control modules were performed on a computer running Inter Core i5-4570 @ 3.2GHz CPU. The 3D reconstruction of unseen objects consists of several steps, including object scanning, reconstruction, stable pose imagination, and robot reorientation. The time duration for each step in a complete experiment trial is listed in Tab. I Overall, our unoptimized code costs an average of 358s, 279s, and 166s on object reconstruction, chair preparation, and seating the bear, respectively.

C. Glossary

1) Parameter Tuning: Our method mainly consists of an object reconstruction module, an imagination module, and a planning and control module. *Parameter tuning* refers to the process of determining and adjusting the parameters in these modules. Specifically, in imagination, parameter tuning aims to ensure that the imagination provides a reliable representation of actions and an accurate evaluation of the resultant configuration. We first tune the dynamic parameters of the simulation environment to decrease the sim-to-real gap. Next, we tune the parameters of imagination setting which includes the scale and initial pose of the agent, and the angular increment of the agent and object. Furthermore, we tune the parameters of the sitting affordance model to optimize the model's ability to correctly identify failure sittings. We use the synthetic dataset in [2] and a real chair to determine the parameters in the imagination module. For the reconstruction module and the planning and control module, we use the real chair to determine their parameters.

2) Perceptibility: Perceptibility refers to the degree to which an object can be perceived by the robot in a certain configuration. For example, when a chair is overturned on a table, the seat and the back are occluded, and the rest , which can be captured by the robot, is perceptible. The perceptibility of the chair in different poses is different. We use the *perceptible difference* to describe this variation in perceptibility.

ACKNOWLEDGMENT

The authors thank Jikai Ye, Xiaoli Wang, and Wanze Li for their helpful discussions, and all the annotators for the human annotations.

REFERENCES

[1] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3762–3769.

[2] H. Wu, D. Misra, and G. S. Chirikjian, "Is that a chair? imagining affordances using simulations of an articulated human body," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7240–7246.