The SFU-Store-Nav 3D Virtual Human Platform for Human-Aware Robotics

Bronwyn Biro, Zhitian Zhang, Mo Chen and Angelica Lim

***Abstract*— Human-robot interactions are challenging to test, and errors in real environments can be costly and dangerous. While 3D simulations aim to provide a safe way to experiment with robot navigation systems, existing simulations fail to capture natural human behavior like inspecting products, wandering, and doubling back. We present a 3D simulation of the SFU-Store-Nav dataset [1], a video and motion capture dataset collected from 108 participants undergoing a shopping experiment and interacting with a Pepper robot. The retail scene was re-created in Blender, and the 3D body shape and pose estimations were combined with motion capture data to create virtual humans that interact with the environment as in the original experiment. This virtual human platform aims to provide a safe, quick, and inexpensive way to test human intent inference and robot navigation systems. We propose this 3D dataset for use in retail-related human navigational intent and human-robot interaction experiments, and include a baseline for human pose forecasting tasks in this new environment.**

I. Introduction

Safe, natural human-robot interactions require an understanding of human motion and intent. Virtual human platforms can be used to experiment with systems that utilize human motion in a quick and inexpensive manner. For example, to see when a retail or domestic robot offers assistance to a human while developing a human intent prediction algorithm, to visualize how a mobile robot navigates a crowd, or to test whether an autonomous vehicle can accurately assess a pedestrian as a risk. However, there is a lack of datasets involving natural, full-body human behavior in simulation. Existing datasets for virtual human motion either lack environmental interaction by overlaying human actions on static backgrounds [2], [3], or model human motion with a standard walk algorithm at a constant speed [4] rather than capturing realistic human behavior such as wandering, inspecting products, and periodically stopping in aisles. To address this, we present a 3D simulation of the SFU-Store-Nav dataset [1], which contains video and motion capture data from people undergoing a shopping simulation and interacting with a Pepper robot [5] when they need assistance. We propose this virtual dataset for visualizing and experimenting with human-robot interactions, human navigational intent and motion prediction experiments, and provide a baseline for human pose forecasting tasks in this new environment.

Authors are with the School of Computing Science, Simon Fraser University, Burnaby, BC Canada. {bbiro, zhitianz, mochen, angelica}@sfu.ca

A. Related Work

Sim4CV [6] is a photo-realistic training and evaluation simulator built on top of the Unreal Engine that integrates physics based cars and animated human actors, allowing for simulation of autonomous driving, flying, and aerial object tracking. Existing datasets involving human motion tend to use motion capture data to generate virtual humans performing a variety of actions and overlay it on various background images [2], [3]. Since the generated avatars move independently from the environment, they do not interact with the environment in an appropriate way, such as walking over a kitchen countertop. HumANav [4] features virtual humans interacting with a virtual office, but use a standard walk algorithm at a constant speed, which fails to capture realistic human behavior such as wandering, inspecting products, and periodically stopping in the aisles. Data shared by Brščić et al. [7] includes pedestrian tracking in a shopping mall, which provides natural human motion but does not offer a virtual 3D simulation.

To address the lack of realistic human behavior in existing 3D simulation platforms, the SFU-Store-Nav 3D Virtual Human Platform captures natural human behavior such as inspecting products, wandering, and doubling back.

II. Proposed Dataset

In this work we present a 3D virtual environment of indoor human navigation experiments. This virtual environment reconstructs the real-life human-robot interaction experiments captured by the SFU-Store-Nav dataset. Our method can be applied to new datasets that contain both video and motion capture data.

A. SFU-Store-Nav Dataset

This dataset contains visual recordings and motion tracking data of human participants, collected through a series of experiments which simulate a shopping scenario. 108 human participants were asked to locate and pick up items from their shopping list and interact with a Pepper robot programmed to help the human participant. The visual data are videos recorded through five cameras placed in each corner and the center of the room. The ground truth human participants' head (x, y) position and orientation in (roll, pitch, yaw) were captured using a Vicon motion capture system. A total of 100 trials corresponding to roughly 1,100 minutes of human movement were selected to reconstruct in simulation.

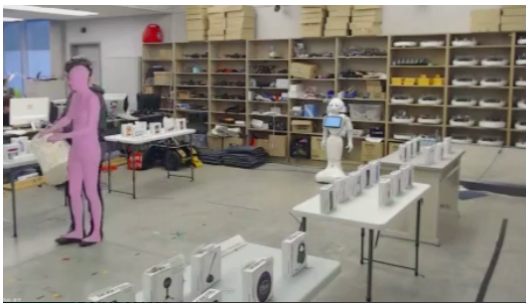
B. 3D Virtual Environment

1) Obtaining pose estimates: To create the virtual environment, we transform each trial from the SFU-Store-Nav dataset into a 3D simulation using Blender. For each experiment, the five videos were first processed using VIBE [8] to obtain the 3D body pose for each frame in the Skinned Multi-Person Linear Model (SMPL) [9] format. We select one full video with the best result for further processing, allowing this method to be used even if there is only video data from a single viewpoint. The result is a collection of body meshes in OBJ format, as well as a mapping of frame number to SMPL pose parameters. There are 72 such parameters, which correspond to the 24 SMPL joints in the 3 dimensions.

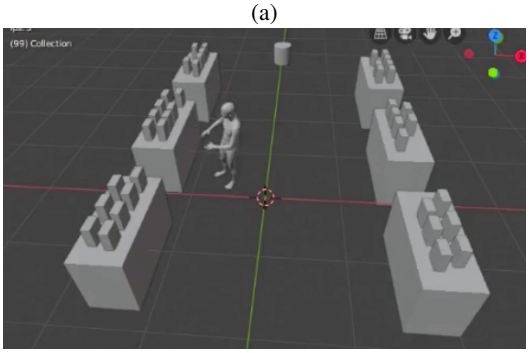
2) Importing poses to Blender: The collections of SMPL body meshes are imported into Blender by using the Stop Motion Obj Blender plugin [10]. The meshes are then transformed into a sequence where each mesh corresponds to a frame in a rendered animation. This animation shows the virtual human going through each pose in a static position within the environment.

3) Applying position and orientation: Lastly, we apply the proper position and orientation from the motion capture data to the animation. We apply the x, y positions from the motion capture data, as well as roll, pitch, and yaw. This allows the virtual human to move throughout the environment. While head-mounted motion capture was used in our study, future work could include exploring vision-based tools to extract position and orientation from video only.

We also rebuilt the indoor scene of the experiments in the 3D simulation. By combining the 3D human body pose, indoor scene, and the motion tracking data of the human, we can replay the SFU-Store-Nav dataset in a simulated environment. Examples are shown in Figure 1. Our code[1] and the 3D virtual environment[2] are publicly available as an extension to the current SFU-Store-Nav dataset.



(a)



(b)

Fig. 1: (a): SMPL body pose and shape estimate predicted by VIBE for an experiment. (b): Our virtual environment in Blender for the same environment. A video demonstration is available at `https://youtube.com/playlist?list=PLYDGbivlPATIZkNaXE75MEkDkQNDt2r1m`

III. Applications

A. Pose Forecasting

As a proof of concept application, we applied our dataset to a pose forecasting task. We implemented a Long Short-Term Memory (LSTM) Variational Autoencoder [11] to forecast future poses. The input to our encoder is a sequence of 3D poses $P_{t-k...t}$ from the lookback period k up until the pose P_t at time t, and we use a probabilistic encoder-decoder to estimate a probability distribution of the future poses $P_{t+1...t+k}$ [12]. As in [13] we use 21 joints, the 24 main SMPL joints but without the pelvis and hand effectors. This allows for compatibility with variations of the SMPL model such as SMPL-X [13] or SMPL-H [14]. Following [15], we use Mean Per Joint Position Error (MPJPE) in meters as our evaluation metric, which measures the Euclidean distance between our generated future poses and the ground truth future poses. To test our method, 10% of frames were randomly selected from each experiment as start times for a sequence, for a total of 6,355 sequences and 38,130 frames. We report a baseline MPJPE of 0.22m per frame when $k = 6$ and 0.24m per frame when $k = 9$, for forecast lengths of approximately two and three seconds respectively.

[1] `https://github.com/bronwynbiro/human_body_prior`
[2] `https://www.rosielab.ca/datasets/sfu-store-nav`

B. Future Applications

Our 3D virtual environment may be beneficial for visualizing and experimenting with human-robot interactions, human navigational intent prediction, human pose and trajectory forecasting, and robot navigation systems. While it may not be suited for end-to-end visual learning without textures, we suggest that this dataset will already be useful for algorithms that work on skeleton representations of humans.

IV. Conclusion

We propose a new 3D simulation that features virtual humans undergoing an indoor shopping scenario and interacting with a Pepper robot. For future directions, we aim to incorporate more detailed 3D human models such as those that construct a natural face, hair, clothing, and texture [16], [17]. We also aim to model deformable objects such as shopping bags, import the Blender objects into Gazebo [18] for use with ROS [19], and explore using long-term human motion forecasting to generate novel sequences. Furthermore, we aim to integrate additional controllers into the 3D humans, to allow them to react naturally when faced with robot interference.

REFERENCES

[1] Z. Zhang, J. Rhim, M. T. Ahmadi, K. Yang, A. Lim, and M. Chen, "Sfu-store-nav: A multimodal dataset for indoor human navigation," *Data in Brief*, vol. 33, p. 106539, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920314219

[2] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] A. Pumarola, J. Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3DPeople: Modeling the Geometry of Dressed Humans," in *International Conference in Computer Vision (ICCV)*, 2019.

[4] V. Tolani, S. Bansal, A. Faust, and C. Tomlin, "Visual navigation among humans with optimal control as a supervisor," *IEEE Robotics and Automation Letters*, vol. 6, pp. 2288–2295, 2021.

[5] "Pepper." [Online]. Available: https://www.softbankrobotics.com/emea/en/pepper

[6] M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, "Sim4cv: A photo-realistic simulator for computer vision applications," *International Journal of Computer Vision*, vol. 126, no. 9, p. 902–919, Mar 2018. [Online]. Available: http://dx.doi.org/10.1007/s11263-018-1073-7

[7] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 522–534, 2013.

[8] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, Oct. 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[10] Justin, "Stop motion obj," May 2021. [Online]. Available: https://github.com/neverhood311/Stop-motion-OBJ

[11] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 843–852.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[13] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, Nov. 2017.

[15] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *European Conference on Computer Vision (ECCV)*, September 2018.

[16] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *International Conference on 3D Vision*, Sep 2018, pp. 98–109.

[17] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[18] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," 2004.

[19] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org