Video Game Exploit: Hacking Game Assets to Learn 3D Human-Object Interactions

Paper Submission ID 639

Abstract

Although 2D human-object interactions have been studied extensively in various forms, the geometrical relationships in 3D human-object interactions remain largely untouched. One major impeding factor is the lack of well-annotated dataset with rich 3D information. In this paper, we introduce a large-scale synthetic dataset SHADE (Synthetic Human Activities with Dynamic Environment) to alleviate such problem. Unlike current 3D datasets that capture real-life human behaviors using motion capture cameras, we collect the dataset by hacking the asset of an open-world gaming platform that has abundant daily activities with human-object interactions. We show that our dataset provides rich and realistic geometrical relationships in human-object interactions that can generalize to real-life human activities. We believe that our dataset opens up new possibilities and challenges for understanding how human interact with the rest of the world.

Introduction

Understanding the geometric relationships in human-object interactions (HOI) is beneficial to many real-life tasks such as robot grasping, surveillance, human activity analysis, and object detection. Although we have seen rapid growth in the analysis of 3D humans and 3D scenes over the past few years, there are very few works that focus on modeling the interaction between human and dynamic objects in 3D.

Notations. In this paper we use the term *static object* for objects that do not move over time, and *dynamic object* for objects whose position or orientation changes over time. In this paper, we only consider objects that move due to HOI and neglect other factors such as gravity.

The difficulty lying behind the challenge of modeling dynamic human-object interactions is mainly two-fold:

Heavy occlusion. When a person interacts with an object, it is very natural for the person to partially, if not completely, occlude it in front of a camera. For smaller objects such as a cup or a pen, the object is very likely to be completely occluded by the interacting person. This creates formidable challenges for detection-based object localization algorithms. (Wei et al. 2013b) argues that we can predict dynamic object location from the 3D skeleton of the interacting human, whose estimation has been widely studied. However, detailed annotation of 3D object locations is exceptionally difficult to acquire in real life, which leads to the next difficulty.

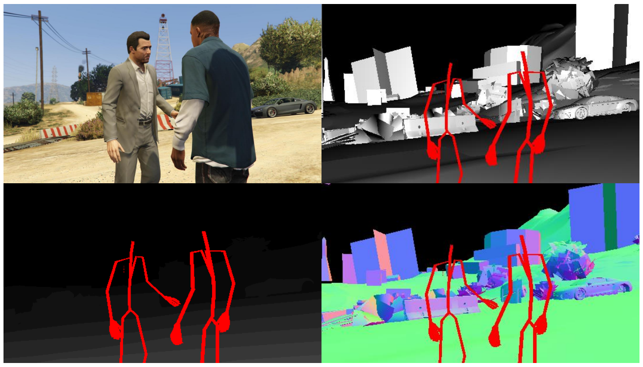

Figure 1: Illustration of synthetic data in SHADE.

Lack of data. Existing dataset are most likely to focus on two aspects separately: dynamic human analysis (Ionescu et al. 2014) or static scene analysis (Song et al. 2017). Although some existing datasets (Savva et al. 2014; Wei et al. 2017) does contain 3D human-object interactions, they lack either the annotation of dynamic object location or annotation of object location in general. (Koppula, Gupta, and Saxena 2013) does provide annotation of dynamic object location but is limited by its data complexity as well as its annotation granularity since it only contains 3D positions of 15 joints and does not provide the 3D geometry ground truth for objects.

In this paper, we present a large-scale dataset SHADE (Synthetic Human Activities with Dynamic Environment) in order to alleviate both difficulties by utilizing the graphics engine in a video game that contains abundant human-object interactions. Our dataset tracks the 3D skeleton of every human accurate to three knuckle joints in each finger and contains real-time 3D position, orientation, and geometry of every object as small as a piece of potato chip. Our experiment reveals three properties of our dataset: i) modeling of human-object interaction provides a significant edge to understanding human behavior; ii) the geometric relationship in human-object interactions can be generalized to real-world human activities; iii) in addition to having more

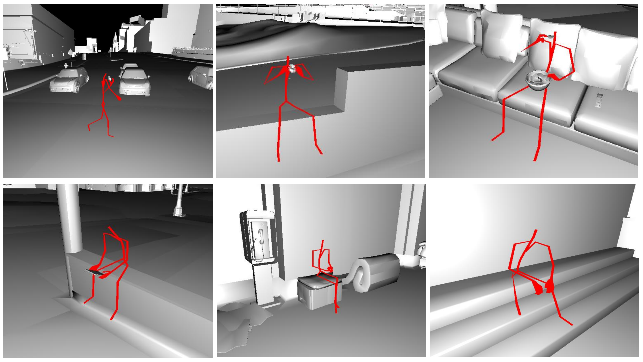

Figure 2: Illustration of variances of the same action. The first row belongs to eat and the second row belongs to sit.

detailed annotation, the human skeleton in our dataset is a complement to other public human pose datasets.

Related Work

Our work is closely related to the following three research streams.

2D/3D HOI recognition. Rather than detecting objects or estimating articulated human pose individually, recognizing human-object interactions (HOIs) requires a deeper and more comprehensive understanding of the mutual spatial structure information and rich semantic relations between human and objects. HOI recognition has gained increasing research interests over the past few years. Earlier methods (Gupta and Davis 2007; Gupta, Kembhavi, and Davis 2009; Yao and Fei-Fei 2010a; 2010b; Yao et al. 2011; Delaitre, Sivic, and Laptev 2011; Desai and Ramanan 2012; Hu et al. 2013) were mainly based on handcrafted features (*e.g.*, color, HOG, and SIFT) with object and human detectors. More recently, with the popularity of deep learning technique in computer vision, various network architectures (Mallya and Lazebnik 2016; Shen et al. 2018; Chao et al. 2018; Gkioxari et al. 2018; Qi et al. 2018) were explored for tackling this task. Some large-scale 2D datasets (Chao et al. 2015; 2018) were also proposed to support the training of deep HOI model.

However, most of the previous attempts focused on HOI recognition in 2D images. Only a few methods (Wei et al. 2013a) were proposed for modeling HOI in 3D scenes. Despite the difficulties brought from the extra dimension, the lack of a large-scale, well-annotated 3D HOI dataset severely restricted the development of 3D HOI recognition. In this work, we propose a large-scale, synthetic 3D dataset for HOI recognition, which is long-time urged in this field. We believe that this dataset would open up new possibilities for moving HOI recognition and analysis into 3D.

Action recognition. There are two main streams in current action recognition methods: appearance-based or skeleton-based.

Similar to HOI recognition methods, researches in appearance-based action recognition have moved from hand crafted features (Willems, Tuytelaars, and Van Gool 2008) to learning deep features with neural networks (Ji et al. 2013; Simonyan and Zisserman 2014; Karpathy et al. 2014). Recently, appearance-based action recognition methods (Zhuang et al. 2017; Heilbron et al. 2017) have seen significant improvement in both classification accuracy and generalization capability by incorporating contextual information.

Skeleton-based methods (Wang and Wang 2017; Yan, Xiong, and Lin 2018), on the other hand, are more robust against appearance and lighting changes since they ignore image features altogether. However, the use of contextual information is very limited in skeleton-based action recognition, largely due to the lack of well-annotated data.

Object Localization.

Object localization has long been a challenging task for computer vision. In 2D object localization, a common practice is to use a sliding window and to run object detection algorithm on each window (Lampert, Blaschko, and Hofmann 2008). This stream naturally extends to convolutional neural networks. Others (Oquab et al. 2015; Tompson et al. 2015) regress heatmaps of object presency on images directly. The recent development of convolutional neural networks have yielded huge leap (He et al. 2017) in 2D object localization by extending and combining both ideas into the region of interest (ROI) operations. In 3D, however, object localization remains challenging due to the cubic growth of data size brought by the extra dimension. (Song and Xiao 2016) extended the sliding window to 2.5D by applying convolutional neural network on RGB-D images. However, such method is sensitive to occlusion.

In addition, many works (Huang et al. 2018; Izadinia, Shan, and Seitz 2017; Lee et al. 2017) have been done to estimate the static scene layout given a 2D image. However, small and dynamic object localization in 3D has yet to be addressed due to the lack of data.

Method

In this section, we will describe the method we use for collecting data from the graphics engine. Figure 3 illustrate the pipeline of our method.

Photo-Realistic Physics-Realistic Synthetic Game Environment

Although human activities involving objects are ubiquitous in daily life, the effort to record such fruitful interaction data to a fine-grained level remains challenging.

Some resort to optical motion capturing systems for target localization, *e.g.*, VICON cameras (Ionescu et al. 2014; Sigal, Balan, and Black 2010). Others make tactile sensors to estimate hand pose during interaction such as tactile gloves (Liu et al. 2017; Edmonds et al. 2017). These approaches require elaborated system set up to serve real-time data recording. In our approach, instead, we build our data acquisition pipeline based on a video game platform – Grand Theft Auto V (GTA V). Unlike other video games which simplify the dynamics of human-object interaction, GTA V is well-known for its richness in photo- and physics-realistic daily activities. In this video game, abundant human-object

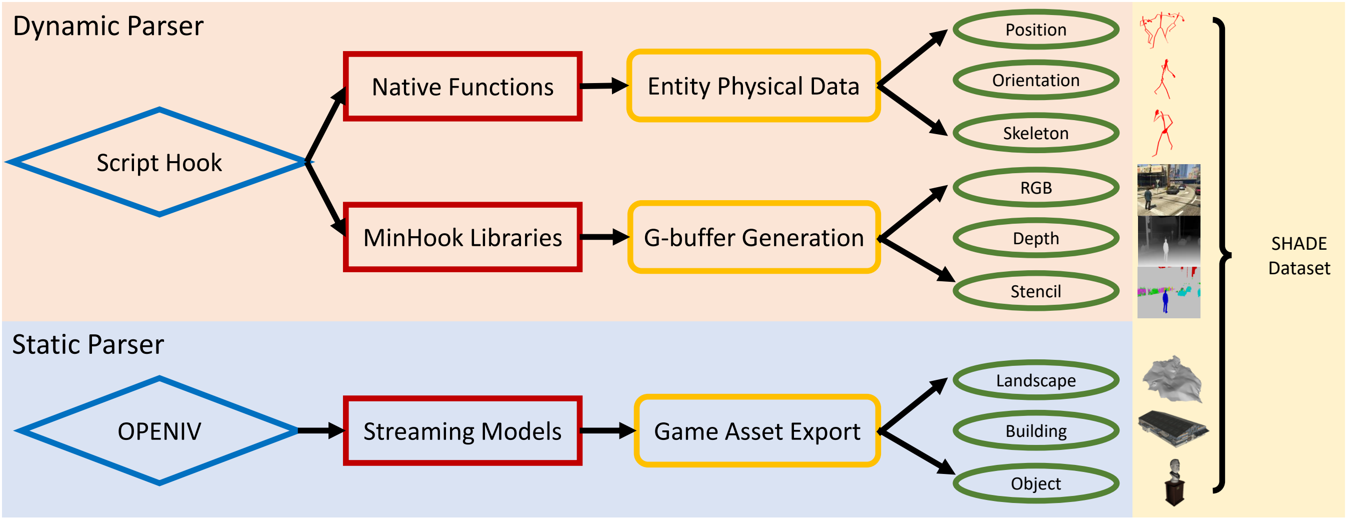

Figure 3: Pipeline of our data collection method.

interaction events are incorporated. For instance, we can see a human agent walking in the street eating a sandwich, and another human agent sitting on a low wall reading from a tablet. In order to obtain the interaction data of agents and objects, we develop a game plugin as the game data parser running parallel with the rendering process.

Game Plugin Design and Characteristics

The development of our data acquisition plugin is based upon the Script Hook library which provides an accessible interface to the GTA V script native functions. The released plugin is portable to the GTA game running environment and can parse the game data in real-time. We characterize the main features of our plugin as the following:

Data Scope. Once the plugin is hooked up inside the game, it is running silently in the back end for data collection. Though our plugin is capable of retrieving data in the area of the whole game map, we limit the data collection range to a fixed radius w.r.t the main character's position for efficiency considerations. In order to collect the different body motion styles featured in different areas in the game map, we periodically teleport the main character to a predefined sequence of locations across the map. In this sense, we guarantee the diversity of collected data.

Collection Method. By making use of the native functions in GTA, we can access the states of gaming agents with our plugin. We collect the data in two means. First, we collect human-object interactions with dynamic objects such as drinking and smoking in real-time. Such interactions are marked in the graphics engine so that the interacting objects are attached to the corresponding agent. Our plugin can detect such attachment and dump the correlating relationships. Second, the human-object interactions with static objects such as sitting on a low wall and climbing over a tree are unmarked in the system. Therefore, we must annotate them offline. The environment data can be dumped from the game asset library beforehand using OPENIV GTA static parser.

Data Formation. Our plugin runs in the background to collect data in each frame. The data collection rate is empirically set to 10Hz so that it does not interfere with the rendering process. In each frame, the raw data incorporates three types of entities in the GTA environment including human agents, objects, and vehicles. For each entity, the plugin captures the real-time physical quantities such as position, orientation, velocity, acceleration, and heading. Besides, for human agents, our plugin also records skeleton data which contains 98 key points, of which 55 are skeletal joints, 21 are facial bone joints, and the rest are control nodes. We also collected the 3D geometry of each object in the form of 3D meshes, which are dumped from the OPENIV GTA static parser mentioned above.

Copyright Issues

Grand Theft Auto V allows non-commercial use of its content as long as certain conditions, such as no spoilers, are met (RockStar-Games 2017). The content of this game have been used in (Richter et al. 2016) for acquiring semantic segmentation annotations for self-driving cars.

Dataset Overview

In this section, we describe the design and composition of our dataset.

Dataset Collection

We adopt two modes in the data collection process: street mode and theater mode. **Street Mode**. We uniformly create 595 grid coordinates across the game map. At each coordinate, we observe and record all human and objects that reside in the graphics engine, regardless of whether it is rendered on the screen. The humans include pedestrians, drivers, business people, construction workers, gangsters, police officers, etc. Although the action space of the observed agents is limited to a predefined collection of activities, each person adopts a different style of body motion according to

their gender, age, occupation, and physique. Therefore, we observe a wide variety of body motion sequences. Figure 2 illustrates the wide variance within two action categories.
Theater Mode. In addition to the constrained set of activities collected from the street mode, we also record human and object dynamics in cutscenes. Cutscenes are CG video clips between game events that are performed by real actors and are perfected by professional artists. The dynamics in cutscenes are more diverse and realistic than those collected in street mode.

Notation. Since in our dataset there are multiple human characters in each time step, one can refer to both a time step in the game engine and a snapshot of a human skeleton as a frame. To avoid miscommunication, we denote each time step in the game engine as a *world-frame* and denote a snapshot of a human character as a *person-frame*.

Action Annotation. We ask volunteers to label human actions to each frame and up to one associated object for each action. For example, if a person is sitting while drinking, our volunteer would label the current frame as (sit, chair), (drink, cup) where 'chair' and 'cup' each refer to a specific object instance in the scene. It is impossible to annotate every frame of our dataset since it contains 902,478 world-frames and on average 32 person-frames in each world-frame. We took our best effort to annotate 609,045 person-frames in the training set and 164,628 person-frames in the testing set. We made sure that we have annotated the actions of every performed activity in our testing set.

Dataset Structure

Training and Testing. We segment our dataset into training and testing set according to the way they are collected. Since the street mode produces varieties of repeating activities and the theater mode produces more diverse variations of the same set of activities but with smaller number of frames, it is natural to assign the street data to the training set and the theater data to the testing set.

Scale. We collected 902,478 world-frames and 29,164,913 person-frames, of which 772,229 person-frames are annotated. On average, each annotated person-frame contains 2.03 action labels and 0.89 interacting objects.

Human Skeleton. For each person, we record the 3D positions of 55 joints including three joints for each finger. Figure 5 illustrates the set of joints in our dataset. In addition to skeletal joints, our dataset contains 21 key points on facial bones for expression and gaze analysis, although we do not provide annotations for expression and gaze.

Object Geometry. We represent the geometry of each object as a 3D mesh accompanied with its translation and rotation in each frame. We use the mesh representation instead of the more popular bounding box representation because it contains much richer information and can support more detailed analysis such as analyzing forces, modeling fine-grained geometric relationships, or modeling the relationship between shape and affordance. We express the rotation of an object in quaternions to avoid the singularity problem in the Euler angle expression.

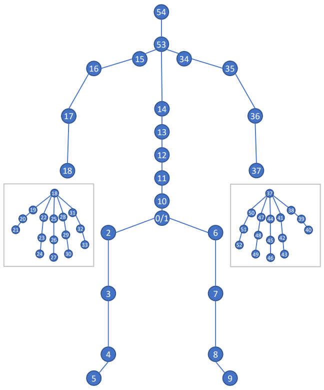

Figure 5: Illustration of human skeleton in SHADE.

	mAP	Top-1 Acc.	Top-3 Acc.
ST-GCN(2018)	0.54	0.35	0.59
2stream(2017)	0.76	0.61	**0.94**
2stream + r	**0.84**	**0.78**	**0.94**

Table 1: Action recognition result.

Comparison with Other Datasets

Existing 3D datasets focus on either human or environment instead of both, with the only exceptions of CAD-60, CAD-120 and, SceneGrok(Sung et al. 2012; Koppula, Gupta, and Saxena 2013; Savva et al. 2014). Table 3 shows the qualitative comparison between our dataset and other 3D datasets. We show that our dataset has richer and more fine-grained annotations than other public datasets.

Experiment

We evaluate our dataset with three tasks: HOI recognition, object localization, and human pose estimation.

HOI Recognition

We run state-of-the-art skeleton-based action recognition models (Yan, Xiong, and Lin 2018; Wang and Wang 2017) on our data and augment the better one with an additional highly coarse contextual feature, richness-of-object, around each joint. Table 1 shows that this simple feature has already provided a significant edge for the state-of-the-art action recognition model.
Richness-of-object. We first uniformly sample point clouds on the surfaces of all objects. Then we compute the number of points within a fixed radius around each joint. We then divide the number by 1000 and clip the result to be between 0 and 1. We append the resulting number to each joint in the human skeleton to reflect the richness of contextual objects around each joint.

Object Localization

We establish two baselines on five common activities in our dataset for the reference of future research. The results are listed in Table 2.

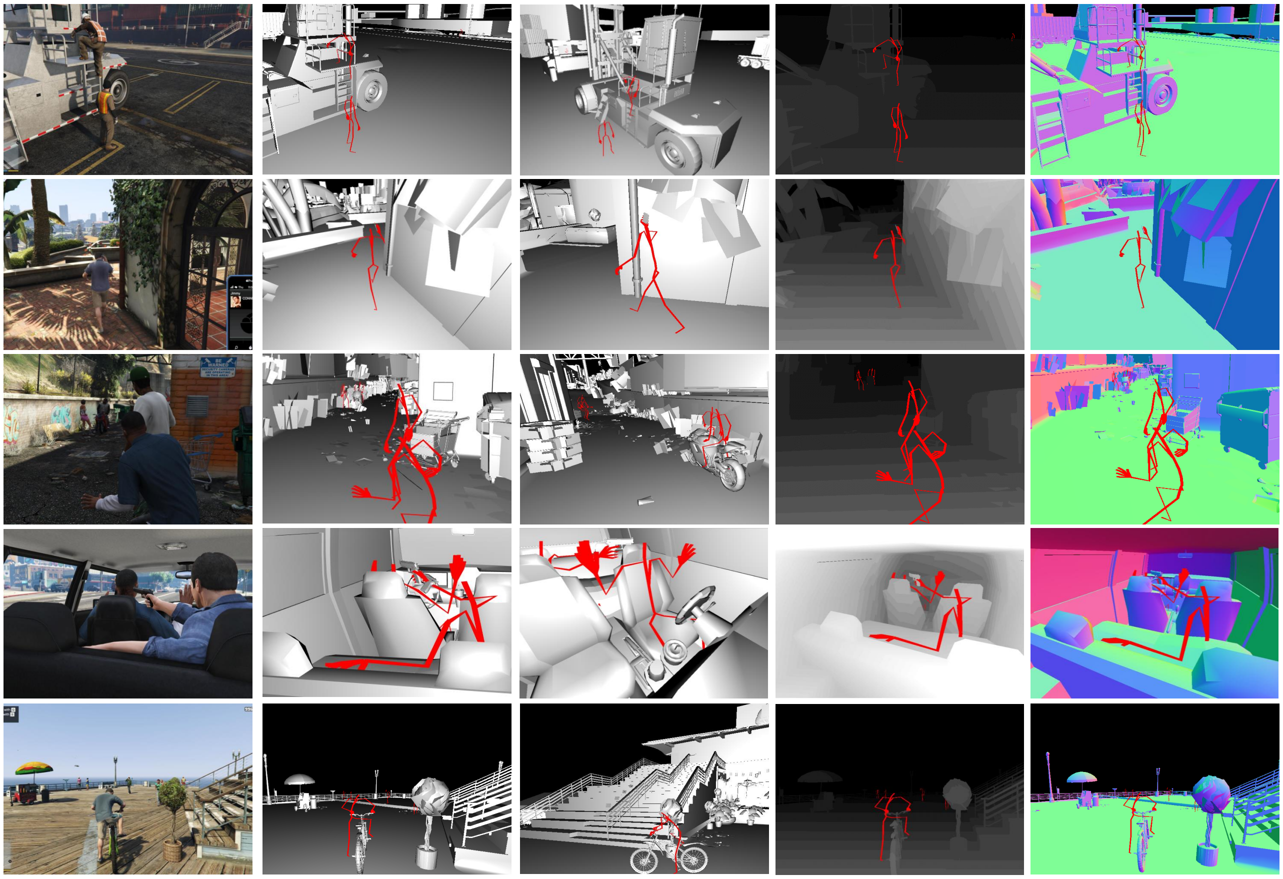

Figure 4: Overview of our SHADE dataset. The six columns are: RGB scene, 3D mesh model, 3D mesh model from novel viewpoint, depth map, surface normal map.

Dataset	# joints	# actions	has object	dynamic	mesh	3D Bbox
HumanEva(Sigal, Balan, and Black 2010)	16	6	No	No	No	No
Human3.6M(Ionescu et al. 2014)	32	16	No	No	No	No
UCLA Multiview(Wei et al. 2017)	20	8	No	No	No	No
MSRA DailyActivity3D(Wang et al. 2012)	20	16	No	No	No	No
SYSU 3DHOI(Hu et al. 2015)	20	12	No	No	No	No
SceneGrok(Savva et al. 2014)	25	7	**Yes**	No	No	**Yes**
CAD-120(Koppula, Gupta, and Saxena 2013)	15	20	**Yes**	**Yes**	No	**Yes**
SunCG(Song et al. 2017)	N/A	N/A	**Yes**	No	**Yes**	**Yes**
SHADE	**55**	**161**	**Yes**	**Yes**	**Yes**	**Yes**

Table 3: Comparison between public 3D datasets

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
H36M	51.8	**56.2**	58.1	59.0	69.5	78.4	55.2	58.1	**74.0**	94.6	62.3	59.1	**65.1**	49.5	**52.4**	62.9
H36M + SHADE	**49.7**	56.6	**57.1**	**58.0**	**67.2**	**77.4**	**54.7**	**57.8**	81.1	**91.5**	**61.0**	**58.5**	65.8	**49.47**	53.2	**62.6**
Protocol #3	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
H36M	65.7	68.8	92.6	79.9	84.5	**100.4**	72.3	88.2	**109.5**	130.8	76.9	81.4	**85.5**	69.1	68.2	84.9
H36M + SHADE	**64.8**	**64.1**	**83.8**	**78.2**	**80.2**	100.5	**67.6**	**84.2**	113.9	**129.1**	**73.5**	**78.0**	85.9	**67.8**	**67.2**	**82.6**

Table 4: Quantitative comparisons of Average Euclidean Distance (in mm) between the estimated pose and the ground-truth on *Human3.6M* under *Protocol #1* and *Protocol #3*. Lower values are better. The best score is marked in **bold**.

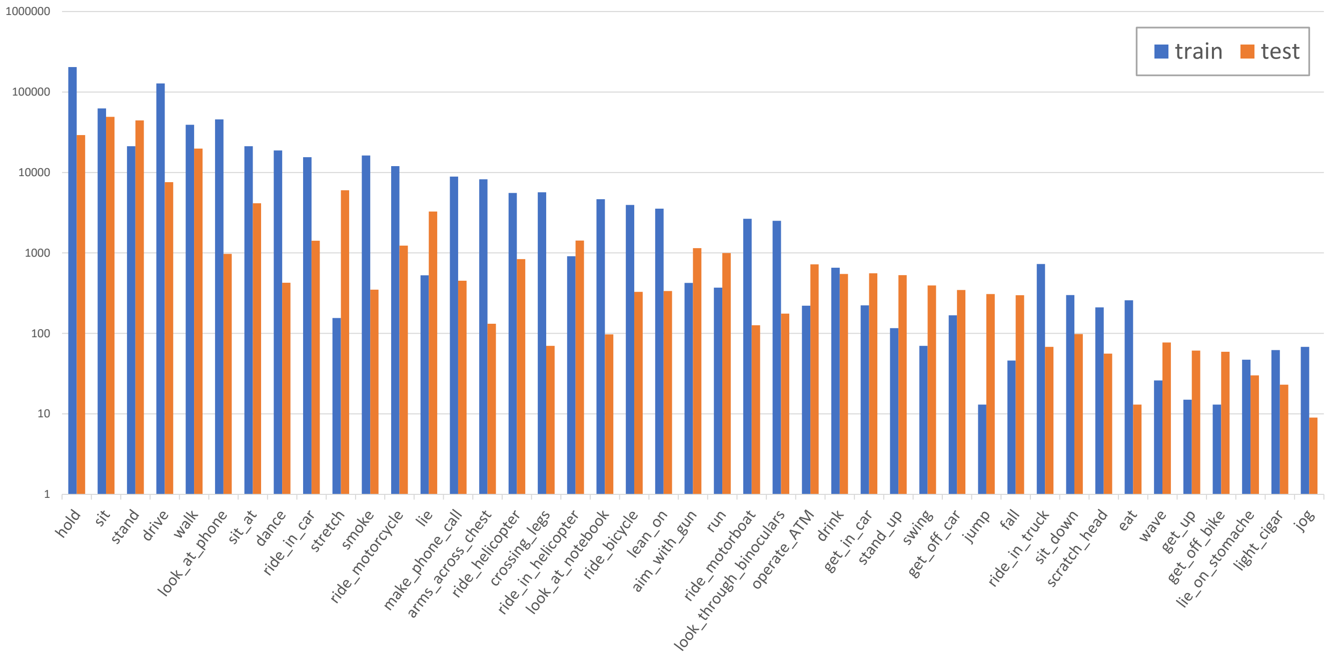

Figure 6: Action frequency histogram in SHADE.

	Smoke	Eat	Drink	Sit	Sit at
KNN	0.08	0.02	0.10	0.37	0.14
DNN(2017)	**0.14**	**0.15**	**0.20**	**0.50**	**0.16**

Table 2: IOU of small object localization.

Referred object. For the first four activities, the referred objects are cigarettes, food, drinks, and chairs respectively. For the last activity, the referred object is the table in front of the sitting person if there exists one.

Baseline methods. For KNN, we normalize each joint coordinate to zero-mean and unit variance and find the first nearest neighbor of the query skeleton in training data. We return the associated bounding box of the nearest neighbor as our prediction result. For DNN, we train a neural network based on the structure proposed in (Martinez et al. 2017) to regress the bounding box coordinates. We evaluate the baseline models on intersection over union (IOU). Notice that the first two activities suffer from extremely low IOU since the referred objects are usually much smaller than other objects, and therefore it is harder for the predicted bounding boxes to intersect with the ground truth ones.

Generalizing to real humans. To show that the geometric relationship learned in our dataset can be generalized to real-world cases, we evaluate the KNN method on a pose chosen from Human3.6M(Ionescu et al. 2014) and show four synthesized objects for eat, sit and sit_at in Figure 8. The selected pose is sitting on a chair and is acting as if she is eating.

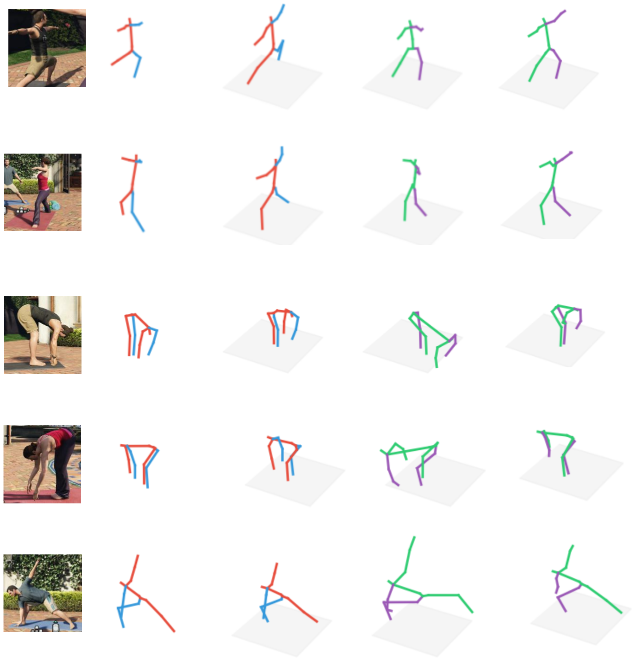

Figure 7: Qualitative result on 3D pose estimation. The first column are querying 2D poses, the second column are ground truth 3D poses, the third column are 3D poses predicted by model trained on H36M, and the fourth column are 3D poses predicted by model trained on both H36M and SHADE.

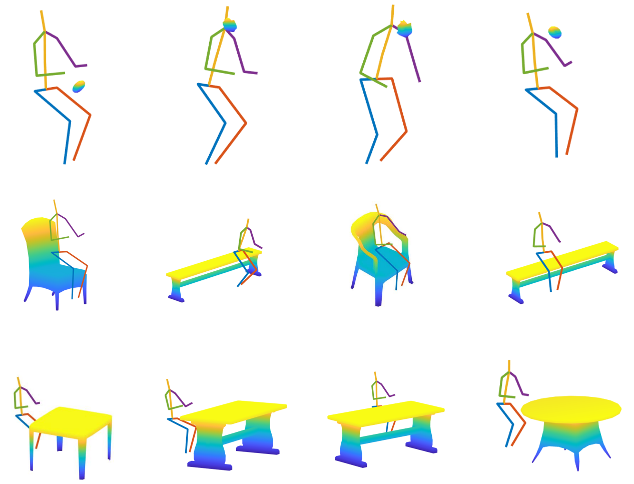

Figure 8: Qualitative result on Human3.6M. The first row synthesizes a bagel or a sandwich for eating, the second row synthesizes a chair or a bench for sitting, and the third row synthesizes a table or a desk for sitting-at. The four columns show four different samples.

Human Pose Estimation

We demonstrate the diversity of our collected human pose in this subsection by training a state-of-the-art 3D pose estimation model (Martinez et al. 2017) on a combination of our dataset and H36M, and compare it with the same model trained solely on the H36M dataset. Table 4 quantitatively shows that the inclusion of our dataset improves the model in most action categories. We further test the two trained models on less common human poses in our testing set, *i.e.*, Yoga poses. Figure 7 qualitatively illustrate that our dataset allows better generalization of the state-of-the-art model than H36M does. We make sure that no Yoga pose or any pose similar to the testing poses are present in the training data.

Conclusion

In this paper, we presented a large scale synthetic dataset SHADE (Synthetic Human Activities with Dynamic Environment). Our dataset is the first that contains rich and fine-grained 3D annotations of human-object interactions. Our experiments show that the human pose in our dataset is a complement to existing human pose datasets and that the geometrical relationship in our dataset can be applied to real-life human behaviors. We believe that this dataset would open up new possibilities in modeling 3D human-object interactions.

References

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *WACV*.

Delaitre, V.; Sivic, J.; and Laptev, I. 2011. Learning person-object interactions for action recognition in still images. In *NIPS*.

Desai, C., and Ramanan, D. 2012. Detecting actions, poses, and objects with relational phraselets. In *ECCV*.

Edmonds, M.; Gao, F.; Xie, X.; Liu, H.; Qi, S.; Zhu, Y.; Rothrock, B.; and Zhu, S.-C. 2017. Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles. In *IROS*.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *CVPR*.

Gupta, A., and Davis, L. S. 2007. Objects in action: An approach for combining action understanding and object perception. In *CVPR*.

Gupta, A.; Kembhavi, A.; and Davis, L. S. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE TPAMI*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*. IEEE.

Heilbron, F. C.; Barrios, W.; Escorcia, V.; and Ghanem, B. 2017. Scc: Semantic context cascade for efficient action detection. In *CVPR*.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; Gong, S.; and Xiang, T. 2013. Recognising human-object interaction via exemplar based modelling. In *ICCV*.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*.

Huang, S.; Qi, S.; Zhu, Y.; Xiao, Y.; Xu, Y.; and Zhu, S.-C. 2018. Holistic 3d scene parsing and reconstruction from a single rgb image. In *CVPR*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI* 36(7):1325–1339.

Izadinia, H.; Shan, Q.; and Seitz, S. M. 2017. Im2cad. In *CVPR*.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3d convolutional neural networks for human action recognition. *IEEE TPAMI* 35(1):221–231.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Koppula, H. S.; Gupta, R.; and Saxena, A. 2013. Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*.

Lampert, C. H.; Blaschko, M. B.; and Hofmann, T. 2008. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*. IEEE.

Lee, C.-Y.; Badrinarayanan, V.; Malisiewicz, T.; and Rabinovich, A. 2017. Roomnet: End-to-end room layout estimation. In *ICCV*. IEEE.

Liu, H.; Xie, X.; Millar, M.; Edmonds, M.; Gao, F.; Zhu, Y.; Santos, V. J.; Rothrock, B.; and Zhu, S.-C. 2017. A glove-based system for studying hand-object manipulation via joint pose and force sensing. In *IROS*.

Mallya, A., and Lazebnik, S. 2016. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *ECCV*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *ECCV*. Springer.

RockStar-Games. 2017. Policy on posting copyrighted rockstar games material.

Savva, M.; Chang, A. X.; Hanrahan, P.; Fisher, M.; and Nießner, M. 2014. Scenegrok: Inferring action maps in 3d environments. *ACM TOG* 33(6):212.

Shen, L.; Yeung, S.; Hoffman, J.; Mori, G.; and Fei-Fei, L. 2018. Scaling human-object interaction recognition through zero-shot learning. In *WACV*.

Sigal, L.; Balan, A. O.; and Black, M. J. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV* 87(1):4–27.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Song, S., and Xiao, J. 2016. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. *CVPR*.

Sung, J.; Ponce, C.; Selman, B.; and Saxena, A. 2012. Unstructured human activity detection from rgbd images. In *ICRA*.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *CVPR*.

Wang, H., and Wang, L. 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *CVPR*.

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*.

Wei, P.; Zhao, Y.; Zheng, N.; and Zhu, S. 2013a. Modeling 4d human-object interactions for event and object recognition. In *ICCV*.

Wei, P.; Zhao, Y.; Zheng, N.; and Zhu, S.-C. 2013b. Modeling 4d human-object interactions for event and object recognition. In *ICCV*.

Wei, P.; Zhao, Y.; Zheng, N.; and Zhu, S.-C. 2017. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE TPAMI* 39(6):1165–1179.

Willems, G.; Tuytelaars, T.; and Van Gool, L. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yao, B., and Fei-Fei, L. 2010a. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*.

Yao, B., and Fei-Fei, L. 2010b. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *ICCV*.

Zhuang, B.; Liu, L.; Shen, C.; and Reid, I. 2017. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*. IEEE.