AUGMENTED ANNOTATIONS: INDOOR DATASET GENERATION WITH AUGMENTED REALITY

Vedant Saran*, James Lin*, Avideh Zakhor

University of California, Berkeley (vedantsaran, james97lin, avz)@berkeley.edu

KEY WORDS: mapping, visualization, tracking, spatial, data, augmented reality

ABSTRACT:

The proliferation of machine learning applied to 3D computer vision tasks such as object detection has heightened the need for large, high-quality datasets of labeled 3D scans for training and testing purposes. Current methods of producing these datasets require first scanning the environment, then transferring the resulting point cloud or mesh to a separate tool for it to be annotated with semantic information, both of which are time consuming processes. In this paper, we introduce *Augmented Annotations*, a novel approach to bounding box data annotation that simultaneously scans and annotates an environment. Leveraging knowledge of the user's position in 3D space during scanning, we use augmented reality (AR) to place persistent digital annotations directly on top of indoor real world objects. We test our system with four human subjects, and demonstrate that this approach can produce annotated 3D data faster than the state-of-the-art. Additionally, we show that Augmented Annotations can also be adapted to automatically produce 2D labeled image data from many viewpoints, a much needed augmentation technique for 2D object detection and recognition. Finally, we release our work to the public as an open-source iPad application designed for efficient 3D data collection.

1. INTRODUCTION

Access to human-labeled data is a necessary component to train supervised models on computer vision tasks. For image data, the ubiquity of smartphones, social media platforms, and sophisticated image processing techniques have made 2D datasets relatively accessible. Annotated 3D data, however, remains relatively scarce. In this paper, we address a bottleneck limiting its availability: the time consuming process of 3D data annotation for the purposes of training and testing.

Current state-of-the-art methods for producing 3D datasets adopt a two-step approach. First, the environment is scanned, often through some sort of tripod-mounted or handheld depth camera system. Next, the scans are uploaded to a server and accessed through a program or web app designed for data annotation. This is inefficient, as it requires two detailed passes over the same environment (once for scanning and once for annotating), and the combined process can be laborious and time consuming.

We point out two key insights that guide our solution to this problem. The first is that real-time Simultaneous Localization and Mapping (SLAM) algorithms have become accurate enough to play a role in generating ground-truth data. The second is that depth cameras have become more accessible in recent years; newer models are cheap and usable with mobile devices.

In this paper we present *Augmented Annotations*, an iOS application that uses a depth sensor to consolidate the scanning and annotation processes for indoor scenes. Users of our app use an iPad to scan the environment while placing virtual bounding boxes that are localized relative to the real world. We show that through our method, users can produce fully-annotated data at a faster rate than through traditional methods.

*Contributed equally

2. RELATED WORKS

Existing methods for annotating large amounts of 3D data utilize crowdsourcing platforms such as Mechanical Turk (Strickland and Stoops, 2018) or oDesk (Wenkart, 2014) to parallelize work. A web application such as SUN RGB-D's annotation tool (Song et al., 2015) presents scanned scenes to each worker, who follow a procedure to create, modify, and label bounding boxes for each object of interest. ScanNet uses a similar web app to produce semantic segmentations (Dai et al., 2017). Such applications are streamlined to minimize the amount of training required.

Augmented Reality (AR) as an interaction paradigm is still in its nascent stages, but nevertheless presents interesting implications for the computer vision community. For example, Alhaija et al. uses AR to generate realistic urban driving datasets (Abu Alhaija et al., 2018). They take real scenes of urban environments and augment them with virtual models of cars and other objects, thus producing endless variants of data from a much smaller library. Industry products like 6D.AI use AR to support persistent annotations for human consumption (6D Development Team, 2017), allowing users to leave virtual, localized notes for themselves or others, amongst other use cases.

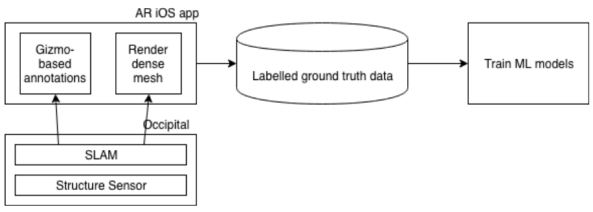

Figure 1: Overview of Augmented Annotations.

3. METHODOLOGY

Our proposed system, shown in Figure 1, uses an iPad connected to an Occipital Structure Sensor, a depth camera and processing

unit that coordinates with the iPad's camera to perform hybrid RGB-D SLAM (Occipital Development Team, 2012). During the scanning process, the sensor provides our application a dense mesh of the environment, which gets rendered over the camera's view of the real world, as shown in Figure 2. A gizmo-style toolkit allows users to insert bounding boxes relative to the world and the mesh; an example of one is shown in Figure 2. Once the process is completed, the position, orientation, extents, and labels of the bounding boxes are exported alongside the mesh itself, where they can then be used as labeled ground truth data for training and testing machine learning models.

Figure 2: Scanned mesh (in white) superimposed on top of camera feed in real-time. A bounding box has been placed around the chair.

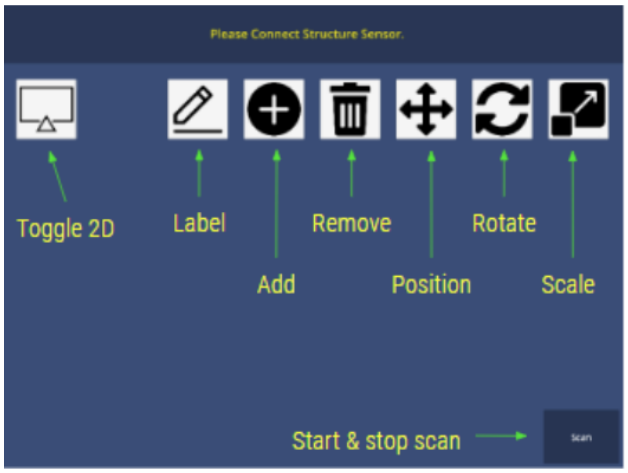

Figure 3: Augmented Annotation's user interface, with labeled tools.

The user interface, shown in Figure 3, supports adding, removing, labeling, and transforming bounding boxes. Tools are accompanied by visual indicators called gizmos, (as shown in Figure 4), which are intuitively operated through taps and drags. Bounding boxes are positioned relative to the real world environment, and maintain their positions even as the iPad moves. This enables users to physically adjust themselves in order to view the scene from a better perspective.

3.1 Automatic 2D Bounding Box Generation

Given a 3D bounding box, our system can generate a 2D bounding box around the object from any perspective, even after the capture is complete. This is done by taking the vertices of the 3D box, projecting them to the camera's image plane, and determining the minimum area rectangle that encapsulates those points as shown in Figure 5. This process requires no additional input from the user besides creating the initial 3D bounding box, and unique images can be generated as fast as the user can move the iPad. This capability is invaluable in generating viewpoint variation of objects in the augmentation of 2D recognition datasets.

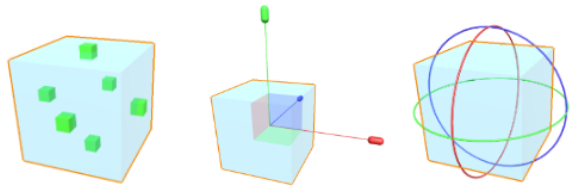

Figure 4: Gizmos for the scale, position, and rotation tools respectively.

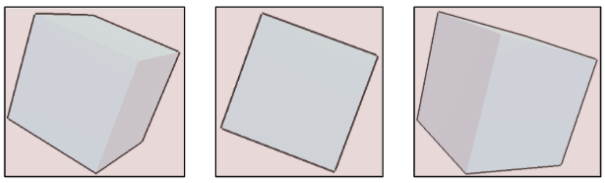

Figure 5: 2D bounding boxes created around a synthetic 3D bounding box. The same 3D box is used in each image, with only the camera perspective changing between images.

4. RESULTS

To evaluate the efficacy of our tool, we compare it to SUN RGB-D, a scanning and annotation system used to generate the eponymous dataset. In SUN's system, the environment is first scanned ahead of time using an RGBD sensor. The mesh of the scan is uploaded to their desktop tool, where workers annotate objects within the scene. For the purposes of comparison, we build a replica of their annotation tool (as shown in Figure 6) and use the Structure Sensor for RGBD capture.

In our experiment, participants are presented with three indoor environments to scan, along with a list of objects in each scene to annotate, as detailed in Table 1. Pictures of the various environments can be seen in Figure 7. Subjects first scan the room with a handheld depth sensor, import it into our replica of SUN's annotation tool, and create labeled bounding boxes around each of the required items. After that, they scan and annotate the scene using our app. The total time required to scan and annotate each environment is compared between the two systems, as seen in Table 2.

Environment	Area (sq. ft.)	No. of Objects
1	189	7
2	71	8
3	85	11

Table 1: The area and number of required objects for each of the environments used in the experiment.

We find that our system completes the scanning and annotation process significantly faster than SUN's approach. In cluttered areas like environment 3, we see significant gains as the scene

	Participants									
	A		B		C		D		Average	
Environment	*SUN*	*AA*	*SUN*	*AA*	*SUN*	*AA*	*SUN*	*AA*	*SUN*	*AA*
1	3:25	2:59	3:28	3:12	4:09	4:25	3:25	3:29	3:36	**3:31**
2	3:56	2:30	4:00	3:40	3:55	3:35	3:57	4:19	3:57	**3:31**
3	6:49	5:16	5:54	4:30	5:27	5:01	5:38	5:01	5:57	**4:57**

Table 2: The times taken for subjects to scan and annotate the environments shown in Figure 7. For *SUN* trials, the time required to scan and time required to annotate are added together. All times are given in minutes and seconds.

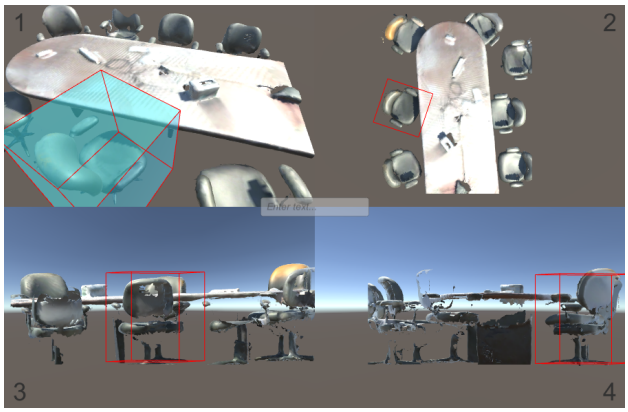

Figure 6: A screenshot from our replica of the SUN RGB-D annotation tool. Workers click to draw out rectangles in the top-down view, then adjust its height in the side views.

Figure 7: The scenes and objects scanned in the experiment. Notice the amount of clutter in environment 3 (bottom).

projects poorly to 2D, making it difficult to distinguish and thus annotate objects using SUN's annotation tool.

We also find variance produced as a result of differing familiarities with the technology. For example, subject A had significantly more experience with AR applications than subject D, and as a result their Augmented Annotations trial times differed greatly. In contrast, all subjects had plenty of experience with mouse-based desktop applications, which can be seen from how the SUN trial times are much more tightly distributed per environment. This suggests that further user studies more rigorous than this informal one would be required to characterize the performance of our system in more depth.

While we had no method of quantitatively evaluating our 2D bounding box generation, examples of it used in physical scans can be found in Figure 8.

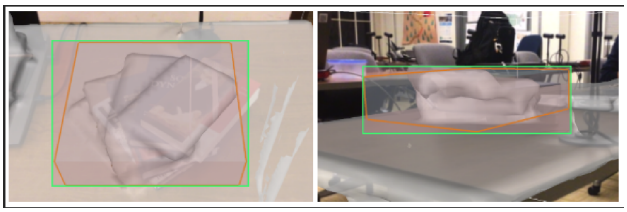

Figure 8: Screenshots of our app generating 2D bounding boxes around a stack of books. The orange wireframe is the 3D bounding box, and the shaded green rectangle is the 2D one.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduce Augmented Annotations, a system for creating 3D datasets that consolidates the scanning and annotation processes. We build an iPad + Structure Sensor app that uses AR technology to enable the real-time creation of bounding boxes relative to the physical world. Our experiments show that our system can outperform traditional methods in generating 3D and 2D bounding box data.

For future work, there are many improvements that could be made to Augmented Annotations to further speed up the process. For instance, bounding box creation would be significantly expedited if the system could make intelligent guesses about the initial placement of the bounding box, such as aligning the bottom of the box to the floor of the mesh. Our general workflow could also be applied to other tasks, such as annotating 3D meshes to train and test semantic segmentation models. All in all, we believe this project represents an untapped potential in the design of intelligent data capture systems.

REFERENCES

6D Development Team, 2017. https://www.6d.ai/.

Abu Alhaija, H., Alhaija, H. A., Mustikovela, S. K., Mescheder, L., Geiger, A. and Rother, C., 2018. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. Int. J. Comput. Vis. 126(9), pp. 961–972.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T. and Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 1, p. 1.

Occipital Development Team, 2012. https://structure.io/.

Song, S., Lichtenberg, S. P. and Xiao, J., 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Strickland, J. C. and Stoops, W. W., 2018. The use of crowdsourcing in addiction science research: Amazon mechanical turk. Exp. Clin. Psychopharmacol.

Wenkart, M., 2014. The Odesk Revolution: Borders are finally a thing of the past. BoD – Books on Demand.