Photogrammetry and VR for Comparing 2D and Immersive Linguistic Data Collection (Student Abstract)

Jacob Rubinstein, Cynthia Matuszek, Don Engel

University of Maryland, Baltimore County – Baltimore, MD 21250
jrubins1@umbc.edu, cmat@umbc.edu, donengel@umbc.edu

Abstract

The overarching goal of this work is to enable the collection of language describing a wide variety of objects viewed in virtual reality. We aim to create full 3D models from a small number of 'keyframe' images of objects found in the publicly available Grounded Language Dataset (GoLD) using photogrammetry. We will then collect linguistic descriptions by placing our models in virtual reality and having volunteers describe them. To evaluate the impact of virtual reality immersion on linguistic descriptions of the objects, we intend to apply contrastive learning to perform *grounded language learning*, then compare the descriptions collected from images (in GoLD) versus our models.

Introduction

The Grounded Language Dataset (GoLD) (Kebe et al. 2021) includes both visual and linguistic data of 47 object types, totaling 207 objects overall. The visual data of the objects in GoLD includes both RGB and depth images from 450 rotational views (see fig. 1), about which 16500 linguistic description were collected, including both textual and spoken descriptions. Kebe et al. (2021) states that in collecting their linguistic descriptions they presented labelers with four key frames of each object (from several hundred in the dataset), making sure to include both standard and non-standard angles of each. This choice was made to avoid a problem of common perspectives recurring in visual datasets leading towards a certain perception or description which does not capture the full nature of the object. We aim to erase this problem completely by taking the data offered by GoLD and using photogrammetry to create high quality, lifelike 3D models for each object. Volunteers will give new linguistic descriptions of these objects by viewing the models in virtual reality, where they will be able to observe our models from all angles. By creating models using photographs from a preexisting dataset instead of collecting a new set of images, we allow the possibility of expanding our research to additional datasets in the future.

Methodology

Data Procurement To store the images from GoLD, Kebe et al. (2021) transformed the depth images into pointclouds which, along with the RGB images, were then recorded in ROS bag files. Bag files list information on 'topics' which are labeled by type of information contained in the bag. We created a virtual machine with Ubuntu 20.04 and installed ROS. We then ran Python code using the $rosbag$ and $rospy$ libraries to obtain RGB images from the '/rgb/image_raw' topic and pointclouds from the '/points2_filtered' topic which we converted back to depth images.

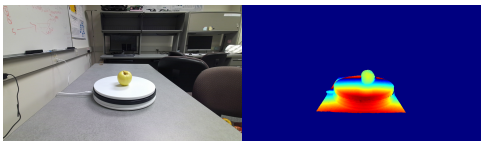

Figure 1: RGB/pointcloud examples from GoLD.

Data Processing Due to constraints of the devices used to collect GoLD's visual data, the amount of RGB and depth images vary per object. In order to perform photogrammetry on the objects, we need matching pairs of RGB and depth images. The ROS bags contain timestamps indicating when each image was taken, which we utilized to find near-simultaneous pairs of images to use. We decided on a time threshold of 0.1 seconds to match pairs of RGB and depth images, as images within that time frame were close enough as to be suitable for photogrammetry. Once this threshold was chosen, we ran an algorithm which culled the non-matching images and labeled the remainders in pairs.

The images in GoLD were collected using a turntable with a lab as a background (fig. 1). This setting was sufficient for generating linguistic descriptions from key frame images, but the detailed background is sub-optimal for photogrammetry, as much of the feature detection will be focused on the background instead of the rotating objects. We will use ImageMagick (The ImageMagick Development Team 2021) masking tools to remove the background in order to improve the quality of our models.

Photogrammetry For the photogrammetry, we will use Open3DGen (Niemirepo, Viitanen, and Vanne 2021), which is "open-source software for reconstructing textured 3D models from RGB-D images." Open3DGen follows a similar pipeline to other photogrammetry software, but best fits our use-case because in addition to using RGB images as input, it also utilizes depth images to speed up computations

and create more accurate models. This software was developed on Ubuntu 21.04, so we created an additional virtual machine using this operating system to run the program.

Description Collection We will generate a 3D scene of a room in which volunteers will be shown models of randomly chosen objects in virtual reality. They will provide one- to two-sentence descriptions, recorded via spoken interaction without leaving VR. Recorded speech will be transcribed using Google's Speech to Text API and the same analysis will be performed: labeling the text samples based on various qualities; rating the accuracy of the speech; and using the Stanford Part-of-Speech Tagger (Toutanova et al. 2003) to count occurrences of each part of speech.

Grounded Language Learning Manifold alignment is a technique in machine learning which allows for different datasets with a nonobservable shared embedding to be projected onto a common "manifold." In the case of GoLD, these datasets are the images and linguistic descriptions of the same objects. The manifold itself can be thought of as a representation of the objects which allows for projection from both "image space" and "linguistic space." If the images are represented as m-dimensional real vectors and the linguistic descriptions are represented as n-dimensional real vectors, then manifold alignment finds the functions ϕ_X and ϕ_Y which project image vectors and linguistic vectors (respectively) into a d-dimensional manifold space:

$$\phi_X : \mathbb{R}^m \longrightarrow \mathbb{R}^d$$

$$\phi_Y : \mathbb{R}^n \longrightarrow \mathbb{R}^d$$

Within this space, the distance between projections shows how well the images and linguistic descriptions match each other, with a smaller distance between projections meaning a greater level of similarity (Nguyen et al. 2021). We will use manifold alignment projections derived from the original GoLD data to project our new linguistic vectors into manifold space, which will show how well our new descriptions match the objects.

Bridging Human Robot Interaction and Computer Vision

The work done by Kebe et al. (2021) to create GoLD and their experiments are deeply rooted in enabling research in the field of human robotic interaction. The objects picked for GoLD were selected specifically for their utility in human-robot learning scenarios, primarily situations occurring in households and domestic situations. Additionally, the manifold alignment model we plan on applying to our collected language samples is very similar to those used in training grounded language models for robots.

Our contribution to GoLD utilizes photogrammetry to create 3D models which will be added to the dataset and used to collect new linguistic descriptions of the objects. The Open3DGen software we will use to generate our models utilizes computer vision techniques in various parts of the photogrammetry pipeline. Niemirepo, Viitanen, and Vanne (2021) indicate that advancements in computer vision techniques have played a large part in the increasing quality of the 3D models produced by photogrammetry.

In non-training HRI situations the environment is not staged, unlike the majority of training scenarios. While of the vision and language information used to train robots for HRI is obtained from standard views of an object, in unstaged environments objects may be seen and described from any perspective. Our use of 3D models placed in virtual reality is much more similar to these unstaged scenarios, as there is no bias towards the common angles of objects found in most visual datasets.

Conclusion

Our work expands on ideas for potential avenues of grounded language research in VR. Using both RGB and depth images from GoLD, we will generate high quality 3D models of objects using photogrammetry. Volunteers will view these models in virtual reality and provide short audio descriptions of the objects. These descriptions will be used as input into the manifold alignment network trained on GoLD data to assess how each aligns to the visual of the object, and to analyze the extent of the effect of VR immersion on linguistic data collection. In addition, our 3D models will be added to GoLD to enable additional future research directions. Our project brings together tools and techniques from computer vision to aid in the acquisition of data which will be used for human robot interaction, allowing us to work towards our objective of diversifying the training data used in that field. Possible extensions of our work could include applying this approach to generate linguistic counterparts to other publicly available asset datasets (e.g., Shrestha 2022).

References

Kebe, G. Y.; Higgins, P.; Jenkins, P.; Darvish, K.; Sachdeva, R.; Barron, R.; Winder, J.; Engel, D.; Raff, E.; Ferraro, F.; et al. 2021. A spoken language dataset of descriptions for speech-based grounded language learning. In *Thirty-fifth Conference on Neural Information Processing Systems*.

Nguyen, A. T.; Richards, L. E.; Kebe, G. Y.; Raff, E.; Darvish, K.; Ferraro, F.; and Matuszek, C. 2021. Practical Cross-Modal Manifold Alignment for Robotic Grounded Language Learning. In *Proc. of the IEEE/CVF CVPR Workshops*, 1613–1622.

Niemirepo, T. T.; Viitanen, M.; and Vanne, J. 2021. Open3DGen: open-source software for reconstructing textured 3D models from RGB-D images. In *Proc. of the 12th ACM Multimedia Systems Conference*, 12–22.

Shrestha, R.; Hu, S.; Gou, M.; Liu, Z.; and Tan, P. 2022. A Real World Dataset for Multi-view 3D Reconstruction. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 56–73. Cham: Springer Nature Switzerland.

The ImageMagick Development Team. 2021. ImageMagick. https://imagemagick.org. Accessed: 2022-12-04.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 173–180. ACL.