Human Factors and Wearable Technologies, Vol. 85, 2023, 30–35
https://doi.org/10.54941/ahfe1003624

A Dataset of Watch and Wristband for Deep Learning Based Multi-View Stereo 3D Model Reconstruction

Bowen Ma, Yi Xiao, Xinyu Guo, and Yuxiang Pang

Artificial Intelligence and Interaction Design, Hunan University, Changsha 410082, China

ABSTRACT

Multi-view stereo (MVS) 3D reconstruction based on deep learning has achieved great success, however, it requires a very high quality and quantity of datasets compared with other computer vision tasks. Current 3D datasets have great limitations in the reconstruction of industrial products, including low accuracy, few types of styles, and few pairwise image models. In this paper, we introduce a new dataset for MVS 3D Model Reconstruction, focusing on the watch wristband category. Better than the existing available open-source watch and wristband dataset, ours contains more than 1k multi-view high-resolution images and high-precision 3D models, covering cartoon, mechanical, vintage, etc. Most importantly, ours can be used directly for deep learning-based MVS 3D reconstruction, because besides three views of real images, we drew line sketches of the three views, and then match them to the high-precision 3D model one by one. At last, we train the MVS network based on deep learning with our dataset as input and supervision. The experiments show that we achieve significant results, and verify the effectiveness of reconstruction in the watch wristband category.

Keywords: Multi-view stereo, 3D data, Watch and wristband, Deep learning

INTRODUCTION

The application of Multi-View Stereo (MVS) 3D reconstruction technology is increasing in various fields, which also means that it is becoming increasingly important to construct datasets that are representative and diverse. Traditional datasets are mainly based on real scenes, such as DTU (Aanæs et al., 2016), ETH3D (Schöps et al., 2017), Tanks and Temples (Knapitsch et al., 2017), and BlendedMVS (Galliani et al., 2015), which provide valuable resources and evaluation tools for MVS 3D reconstruction. However, these datasets often have many limitations, such as excessive scene noise, insufficient focus on object categories, and a complex construction process. Therefore, it is necessary to construct a large-scale dataset with controllable lighting environments, convenient data acquisition, and a focus on specific categories, so that it can more comprehensively support product design reconstruction and meet the needs of different research directions.

Different from previous datasets, this article constructs a dataset based on virtual scenes, which contains hundreds of items of watch and bracelet

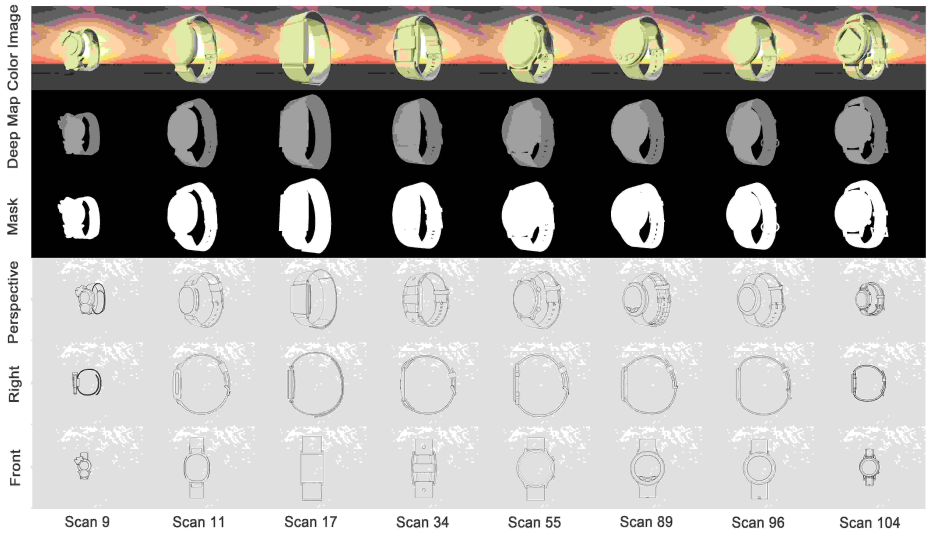

Figure 1: Partial data in our dataset, including color image, depth map, mask, and line drawings from three views.

categories, as shown in Figure 1. We provide a complete image dataset, depth map, and line drawings from three perspectives. Compared with traditional datasets constructed based on real scenes, the virtual scene-based watch and bracelet dataset constructed in this paper can provide higher data control and generation effects, while effectively avoiding various problems in real scenes, such as weather conditions, lighting conditions, noise, and motion blur. In addition, our virtual scene-based dataset construction process is also very friendly to student groups and small teams, which can reduce costs and more flexibly control parameters in the scene, such as camera position, lighting conditions, and object materials, to meet different research needs.

The category of the dataset constructed in our paper has a wide range of applications in intelligent wearable devices and medical equipment. In addition to effectively solving the problem of missing data on watch and bracelet category items in practical application scenarios, it will also provide valuable resources and evaluation tools for related research and applications. Furthermore, our dataset can be widely applied in fields such as MVS 3D reconstruction, computer vision, and machine learning, with great practical value and promotional significance for related research. The sample dataset can be downloaded for free at https://pan.baidu.com/s/1b7N1LgYTM1hM0A9z4-4Aw?pwd = 1234. If you want to obtain the complete dataset, please contact us.

RELATED WORKS

Previous research has investigated various methods for reconstructing 3D objects from multiple views of images, with the most popular method being the use of Structure from Motion (SfM) and Multi-View Stereo (MVS) techniques to generate 3D models from a set of 2D images. The DTU dataset

(Aanæs et al., 2016), the Tanks and Temples dataset (Knapitsch et al., 2017), and the BlendedMVS dataset (Galliani et al., 2015) are the most widely used datasets in the field of MVS 3D reconstruction. As datasets constructed from real-world scenes, they possess authenticity and diversity, which can largely reflect the requirements and challenges of practical applications. However, these datasets have some drawbacks, such as limited size, different scene features and conditions, and varying levels of tests for algorithm robustness and adaptability, making it difficult to comprehensively evaluate the performance of algorithms. Additionally, datasets constructed from real-world scenes are difficult to finely annotate and control due to restrictions imposed by real conditions and scene features, leading to uncertainty in the quality and repeatability of the data.

In addition to datasets constructed from real-world scenes, in recent years, some datasets based on virtual scenes in other fields have emerged. Specifically, the ScanRefer dataset (Dai et al., 2020) contains 200 scenes, each with multiple tasks, such as object localization and relationship inference. The advantage of this dataset is that it can control various parameters in the scene, thereby allowing for systematic evaluation and comparison of different scenes. Additionally, the iGibson dataset (Xia et al., 2020) is another indoor scene dataset built using the Unity engine. It contains over 100 indoor scenes, covering different scene types and sizes, and is closely related to the field of robotics, making it suitable for research in areas such as robot intelligence and autonomous navigation.

However, there is currently no representative virtual scene dataset for MVS 3D reconstruction, which is also the purpose of our dataset construction. Compared to datasets constructed from real-world scenes, our virtual scene dataset has the advantages of large data scale, high annotation accuracy, and strong data controllability, which can provide more standardized, unified, and repeatable data support. At the same time, our dataset constructed from virtual scenes also has features such as scene and material customization, allowing researchers to fully control the environmental conditions and object attributes. Additionally, our dataset focuses specifically on the category of watches and bracelets, which have not been widely studied in previous MVS datasets.

DATA

The Multi-View Stereo (MVS) algorithm can accurately reconstruct 3D models, but various factors in real environments can still hinder data acquisition. To address this problem, we systematically changed the camera positions, scenes, and lighting in a virtual environment. We captured the same 44 positions of hundreds of watch and bracelet models under various backgrounds and lighting conditions, involving camera and lens settings, calibration procedures, and post-processing steps to ensure high-quality and accurate image and depth information for 3D reconstruction.

Image Acquisition & Depth Map Acquisition

Our data acquisition was set up based on Aanæs et al., (2016) to ensure the accuracy and consistency of depth maps. We used a fixed intrinsic parameter

set with a physically calibrated virtual camera, including focal length, image sensor size, and lens distortion coefficients, which are crucial for accurately converting pixel coordinates into real-world coordinates. Additionally, to further improve the accuracy of the depth map, we conducted camera calibration using a checkerboard pattern. This involved capturing several images of the checkerboard in different directions and distances and then estimating the camera's intrinsic and extrinsic parameters using calibration algorithms. We manually inspected the calibration results and improved them using least-squares optimization methods to ensure maximum accuracy. To facilitate obtaining consistent and reliable depth maps, we enabled the depth buffer in the virtual camera's perspective rendering of the scene, which allowed us to capture the depth information of each pixel in the image and save it as a grayscale image file. We used the EXR format(Pike et al., 2003), which allowed us to save depth information as 32-bit floating-point values, maintaining a high dynamic range and minimizing quantization errors.

Position Choice & Lighting Conditions

During the capture of color images and depth maps, we encountered some limitations and challenges. First, the selection of shooting positions for the scene needs to consider factors such as the shooting angle and uniformity of model size. Due to the high position repetition, and the need to accurately obtain camera parameter information through Colmap (Schonberger & Frahm, 2016), we conducted fine design and adjustments in the program and pre-defined accurate positioning by filtering and predefining through modeling software. The 44 shooting positions were placed on a sphere with a radius 1.5 times the minimum bounding sphere of the model body, and a distance of about 0.5 times the radius from the model surface. The scene center is shown in Figure 2. We obtained batches of color images and depth maps from the same positions of hundreds of models in the dataset, which would be difficult to achieve quickly and accurately in general real scenes.

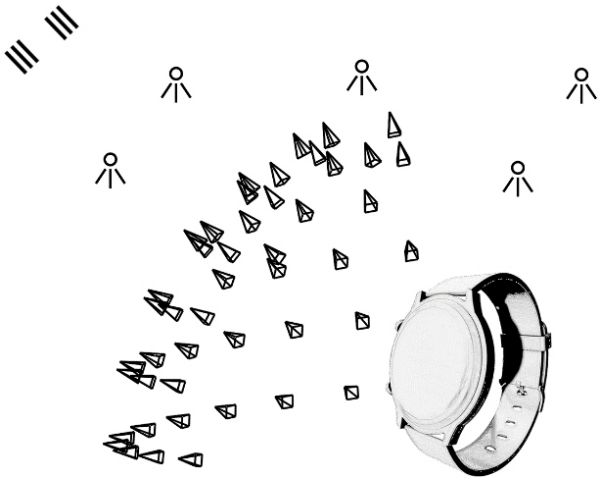

Figure 2: Camera positions on a sphere, 5 point light sources, and 2 parallel lights.

The impact of lighting conditions on the accuracy of color images and depth maps. Shadows, specular reflections, and other lighting artifacts can introduce errors and inconsistencies in the depth information, leading to inaccurate reconstruction. To alleviate this problem, we carefully controlled the lighting conditions, placing a point light source at each of the four corners and the center of a rectangular plane above the scene and adding two parallel lights to ensure uniform illumination of the entire scene. Additionally, we covered all models with neutral soft materials to reduce the impact of lighting artifacts on the images and make them more consistent and reliable. We also used HDR rendering technology in Unity to help us better simulate real-world lighting conditions for more accurate capture of color and depth information.

EXPERIMENTS

Our experiment was conducted on an Ubuntu 18.04 operating system, using 4 NVIDIA RTX 3090 GPUs. We modified the depth acquisition process by accurately obtaining depth values from EXR files, based on the PyTorch implementation of the MVSNet network (Yao et al., 2018) by Guo and Li (2018). We successfully trained our model and achieved a loss of less than 4.9 in just under two hours of training.

Due to the high memory and computational requirements of higher-resolution depth maps, it is not feasible for large-scale datasets. To address this issue, we carefully balanced the resolution and quality of the depth maps with the available hardware resources, setting the resolution of the color images to 1200x1600 pixels with 8-bit RGB color, which was reduced to 480x640 during training, while the depth maps were saved in 128x160 size. In addition, we used image processing techniques such as median filtering and bilateral filtering to post-process the depth maps. This helped to reduce inconsistencies in the depth maps and improve the accuracy of the 3D reconstruction.

Figure 3 show the progression of our loss during training, where we can see that our training loss decreased smoothly alongside the rapid decrease in

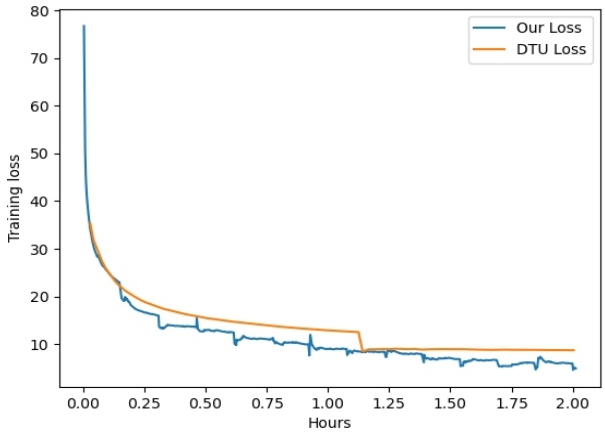

Figure 3: Training loss on DTU dataset and our dataset.

loss on the DTU dataset. Additionally, since our dataset primarily focuses on watches and bracelets, we achieved slightly lower losses than DTU for the same batch, which further demonstrates the effectiveness of our dataset.

CONCLUSION

This paper presents a cost-effective and flexible approach to providing high-quality, consistent, and reliable data for MVS 3D reconstruction research by constructing a large-scale dataset in a virtual environment. Compared to traditional physical capture methods, data acquisition in virtual environments offers higher control over camera parameters and lighting conditions, while reducing errors and inconsistencies in depth maps caused by lighting and object surface reflections. Additionally, data acquisition in a virtual environment can reduce time and cost, making the construction of large-scale datasets more feasible and practical. Finally, our experimental results demonstrate the effectiveness and usefulness of our dataset.

REFERENCES

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., & Dahl, A. B. (2016). Large-scale data for multiple-view stereopsis. International Journal of Computer Vision, 120, 153–168.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017 (pp.5828–5839).

Galliani, S., Sattler, T., Schindler, K., & Pollefeys, M. (2015). Accurate dense and sparsified large-scale MVS with guarantees. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1547-1555).

Guo, X., & Li, H. (2018). MVSNet_pytorch. GitHub. Retrieved from https://github.com/xy-guo/MVSNet_pytorch.

Knapitsch, A., Frahm, J. M., & Pollefeys, M. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5306-5315).

Pike, G., McCann, J., & Veach, E. (2003). The OpenEXR image file format. ACM Transactions on Graphics, 22(3), 684–689.

Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp.4104–4113).

Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., & Frahm, J. M. (2017). A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3260-3269).

Xia, F., Gao, R., Zhang, A., Fei-Fei, L., & Savarese, S. (2020). iGibson: A Simulation Environment for Interactive and Realistic Robotics Vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), (pp.5938–5947).

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European conference on computer vision (ECCV) (pp.767–783).