

이미지 기반 모델러: 효율적인 3D 모델 재구성을 위한 6-카메라 턴테이블 시스템

이효종¹, 딜루샤 드실바¹¹(주)에이아이테크

hlee@jbnu.ac.kr, dmd681@gmail.com

Image-based Modeler: A Six-camera Turntable System for Efficient 3D Object Reconstruction

Hyo Jong Lee¹*, D M De Silva¹¹AI Tech, Inc.

요 약

3차원 모델링의 출현은 전자 상거래와 제품 디자인에서 문화유산 보존에 이르기까지 다양한 산업에 혁신을 가져왔다. 그러나 복잡한 물체의 고충실도 3D 모델을 캡처하는 것은 여전히 도전 과제로 남아 있다. 본 논문은 물리적 객체의 정밀한 디지털 복제를 효율적으로 생성하도록 설계된 제어 환경 시스템인 이미지 기반 모델러를 소개한다. 설정은 자동 회전 테이블과 통합된 6개의 동기화된 카메라를 통해 다각도의 이미지를 원활하게 캡처할 수 있도록 한다. 이러한 이미지는 고급 구조에서 모션 및 다중 뷰 스테레오 알고리즘을 사용하여 처리되어 상세한 3D 모델을 생성한다. 다양한 질감, 크기 및 기하학적 복잡성을 가진 물체를 포함한 일련의 실험을 통해 회전 비전의 성능을 평가하였다. 결과는 모델 정확도의 향상, 재구성 시간의 단축, 전통적인 정적 다중 카메라 설정에 비해 향상된 표면 세부 사항을 보여준다. 이 연구의 응용은 광범위하며, 빠르고 고품질의 3D 객체 디지털화를 요구하는 분야에 유익하다.

1. Introduction

The proliferation of three-dimensional (3D) modeling across industries, including e-commerce, product design, and cultural heritage preservation, has underscored the need for efficient and precise object digitization techniques. Traditional methods often fall short, either sacrificing detail for speed or vice versa, particularly when confronted with objects of complex geometries or varied textures. In response, this study introduces and evaluates Rotating Vision, a novel, controlled-environment system designed to bridge this gap.

Rotating Vision integrates six synchronized cameras with an automated turntable, capturing multi-angle images that are subsequently processed using advanced Structure from Motion (SfM) and Multi-View Stereo (MVS) algorithms. This synergy enables the rapid generation of detailed, high-fidelity 3D models. To

comprehensively assess Rotating Vision's performance, we conducted a mixed-methods evaluation, combining technical metrics with subjective feedback from professionals in relevant fields. A structured survey presented participants with 3D models of diverse objects, eliciting insights into model accuracy, surface detail fidelity, and overall system efficacy compared to traditional static multi-camera setups. This paper presents the findings of our investigation, discussing the implications for enhanced object digitization and the broader applications of Rotating Vision in driving innovation across disciplines.



(Fig. 1) Object and its 3D Reconstruction(right)

* 본 연구는 중소벤처기업부의 창업성장기술개발사업의 지원에 의한 연구임(RS-2023-00264181)

2. Related works

Reconstructing 3D objects from multi-view images is a well-explored field with far-reaching applications. However, achieving accurate reconstructions typically relies on two key inputs beyond multi-view imagery: precise object masks to segregate the object of interest from its background, and often, manually annotated geometric constraints.

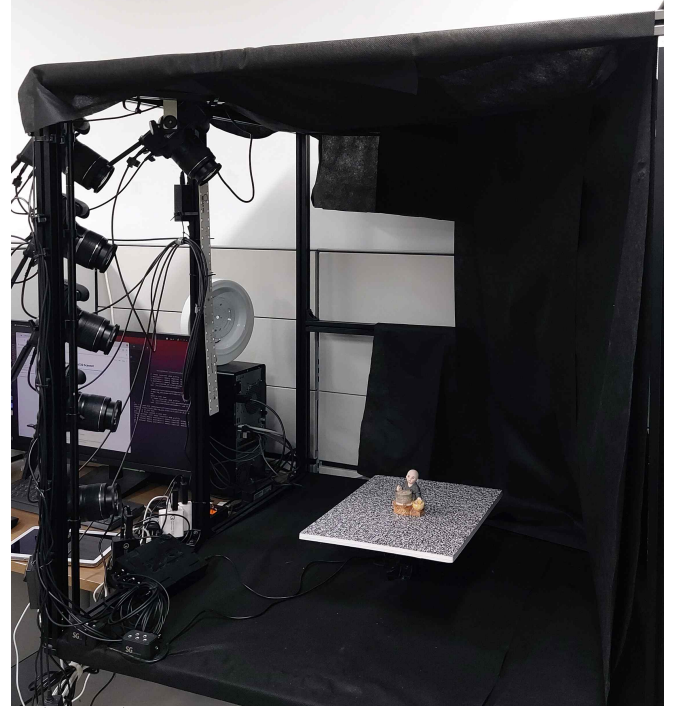
Traditional Multi-View Stereo (MVS) approaches [1, 2] generate background-free models by first estimating depth maps for each frame, then merging them within predefined object boundaries. Recent advances in neural reconstruction techniques, leveraging differentiable renderers and scene representations, have shown significant promise. While surface-rendering methods [3, 4] eliminate the need for 3D supervision, they still depend on object masks as surrogate geometric guides. More recent volume-rendering-based methods [5, 6, 7] enable mask-free training but require supervised object masks to ensure background-free outputs.

3. Methodology

In accordance with [8], point-wise features are extracted by aggregating multi-view 2D DINO features [9], yielding semantically rich outputs, evident in projected colors. The foreground object is segmented from the SfM point cloud utilizing a lightweight 3D Transformer, with point-wise features and a global feature being inputted to predict point-wise labels. Outcome is a 3D bounding box for the object, and optionally a ground plane, are estimated from the decomposed point cloud.

We built a cubical frame structure, equipped with six cameras positioned linearly along its vertical axis. The interior of the cube was illuminated by integrated lighting to ensure optimal image clarity. At the device's core lay a motorized turntable, designed to rotate objects in a controlled manner. To capture comprehensive visual data, objects of interest were placed on the turntable and subjected to an automated photography protocol, wherein the six cameras

synchronously captured images at discrete rotational intervals. The resultant multi-angle image datasets from all six cameras were subsequently aggregated for 3D reconstruction processing, facilitating the generation of high-fidelity, three-dimensional object models. The device structure is shown in Figure 2.



(Fig. 2) Proposed system

4. Experiments

In our experimental setup, a curated set of objects was photographed using the custom-built device equipped with six cameras. Each camera captured a series of images, with the total number of photos per object calculated as $\text{Total Images} = 6 * n$, where n denotes the number of captures per camera. Specifically, we set $n = 20$, resulting in a comprehensive dataset of 120 images per object. To ensure comprehensive coverage, the turntable rotated by a fixed angle after each photo capture. Given a full 360-degree rotation, the incremental rotation angle θ after each capture is,

$$\theta = 360^\circ / n \quad (1)$$

the turntable rotated by a fixed incremental angle of $\theta = 18^\circ$ after each photo capture, facilitating a full 360-degree rotation. Following 3D reconstruction, we calculated a Scaling Factor. To accurately scale the reconstructed 3D

models, we measured the distance between two known points on both the real object (D_{real}) and its constructed 3D model (D_{model}). The Scaling Factor (SF) is calculated as,

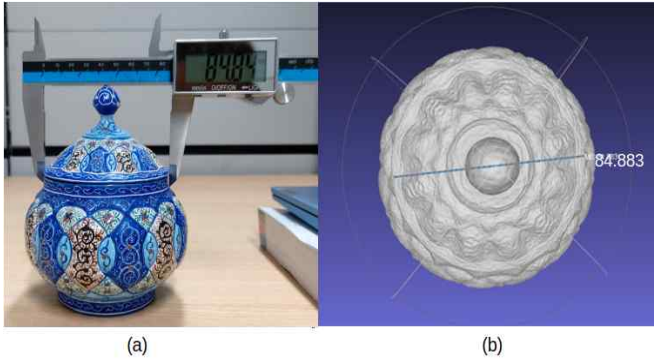
$$SF = D_{real} / D_{model} \quad (2)$$

The accuracy of the reconstructed models was evaluated by measuring the length between two known points on the real object and its 3D reconstruction. The Error is computed using,

$$Error(\%) = ((D_{real} - D_{model}) / D_{real}) \times 100 \quad (3)$$






5. Results

To demonstrate results in a more clear way we show the results for five objects. First we compare point to point distance of the real and constructed for one object (Figure 3). The constructed object is scaled with SF found by (2). Table 1 shows the results with errors calculated by (3) for five objects.



(Fig. 3)(a)Real object measurement(84.84)
(b)Reconstructed model measurement (84.88)

<Table 1> Results for five models

Object	Model	Avg error(%)
Monk		0.09
Green box		0.17
Pot		0.01
Cube		0.01
Plug		0.03

Our experimental evaluation yielded promising results, demonstrating the effectiveness of our multi-camera setup in reconstructing accurate 3D models. With an average Error (%) of 0.0618, our approach successfully captured the geometric details of the objects under study. Notably, the application of the calculated Scaling Factor (SF) ensured a significant improvement in dimensional accuracy.

6. Conclusion

In this paper, we presented a novel multi-camera setup for 3D object reconstruction, leveraging the synergy of six synchronized cameras to capture comprehensive visual data. Through an exhaustive evaluation involving five diverse objects, our approach demonstrated remarkable accuracy with an average error of 0.0618%. The effectiveness of our method in accurately scaling reconstructed models using the proposed Scaling Factor further underscores its practical viability. While our results are promising, future work will focus on exploring deep learning-based enhancements to further refine reconstruction accuracy and investigating the applicability of our setup in dynamic environments. Nonetheless, this research firmly establishes a foundation for the development of more sophisticated, multi-camera 3D reconstruction systems, holding great promise for fields such as computer vision, robotics, and beyond.

Acknowledgement

This work was supported by the Starting Growth Technological R&D program (2023-RS-00264181) funded by the Ministry of SMEs and Startups(MSS, Korea)

References

- [1] Johannes L. Schonberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In ECCV. 2016.

- [2] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In ECCV, 2018.
- [3] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In CVPR, 2020.
- [4] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In NeuIPS, 2020.
- [5] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In ICCV, 2021.
- [6] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In NeurIPS, 2021.
- [7] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In NeurIPS, 2021.
- [8] Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun Bao, Xiaowei Zhou. AutoRecon: Automated 3D Object Discovery and Reconstruction. CVPR 2023.
- [9] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In CVPR, 2021.