Voxel Carving for 3D Human Reconstruction

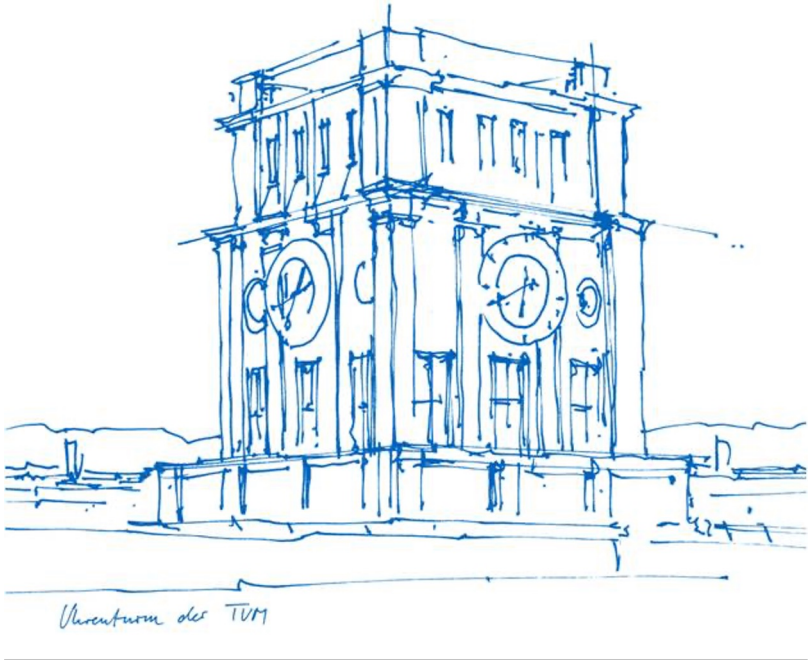
Group 3

Maximilian Anzinger, Alexander Fuchs,

Georg Henning, Nils Keunecke

Technical University of Munich

Munich, 05. August 2022



Overview

- Motivation
- Related Work
- Voxel Carving Pipeline
  - Pose Estimation
  - Image Segmentation
  - Voxel Carving
  - Color Reconstruction
  - Postprocessing TODO -> move later (problems)?
  - Model Conversion: Marching Cubes
- Results
  - Performance TODO -> move to voxel carving?
  - Human Model
- Discussion
- Future Work

Voxel Carving for 3D Human Reconstruction

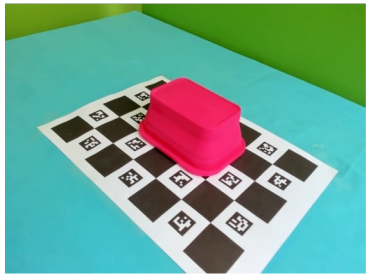
Group 3

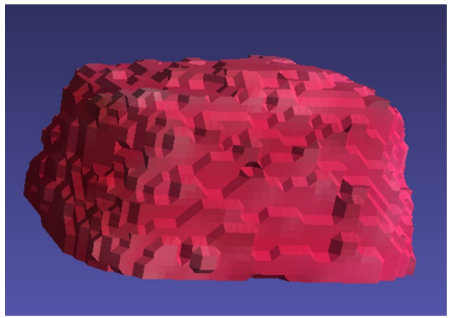
Maximilian Anzinger, Alexander Fuchs,

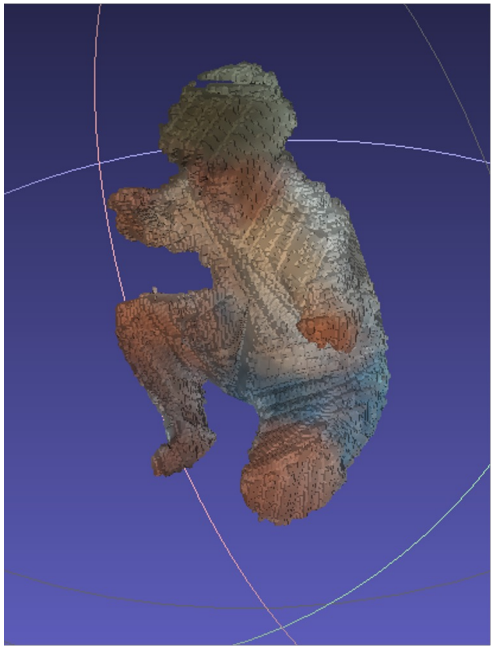
Georg Henning, Nils Keunecke

Technical University of Munich

Munich, 05. August 2022







Motivation & Related Work

Structure from Motion (e.g. Colmap[1])

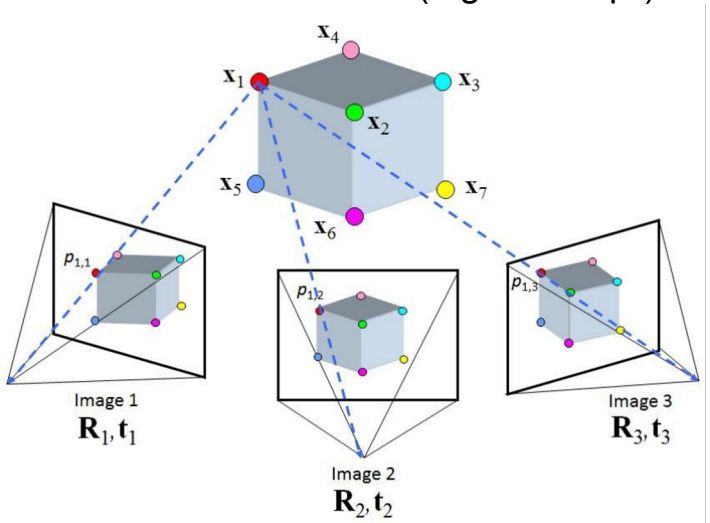


Image source: *Yilmaz, Ozgur & Karakus, Fatih. (2013). "Stereo and kinect fusion for continuous 3D reconstruction and visual odometry"*

[1] *J. L. Schönberger and J. Frahm, "Structure-from-Motion Revisited"*

Motivation & Related Work

Structure from Motion (e.g. Colmap[1])

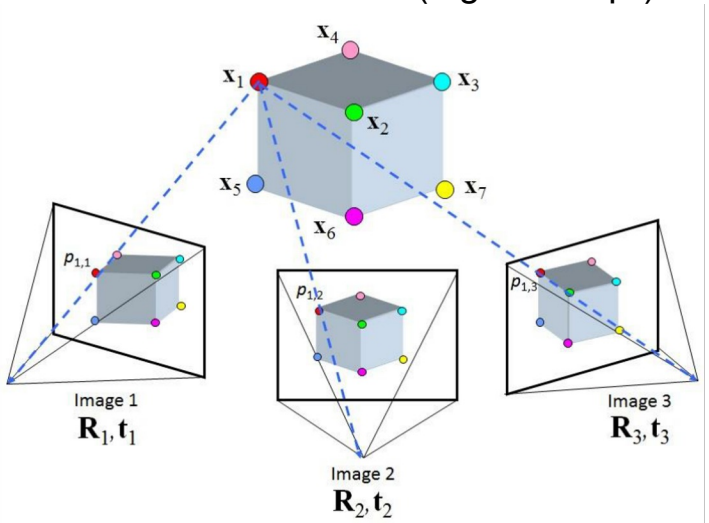


Image Source: *Yilmaz, Ozgur & Karakus, Fatih. (2013). "Stereo and kinect fusion for continuous 3D reconstruction and visual odometry"*

Voxel Carving[2]

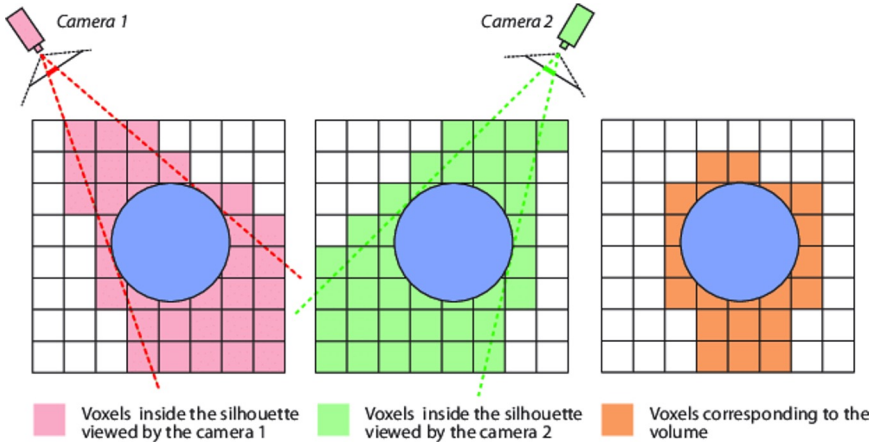
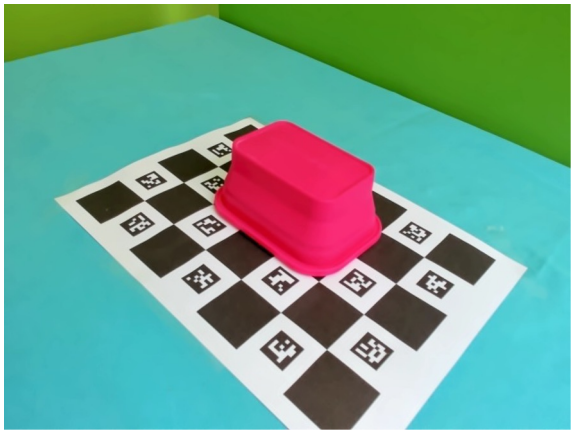


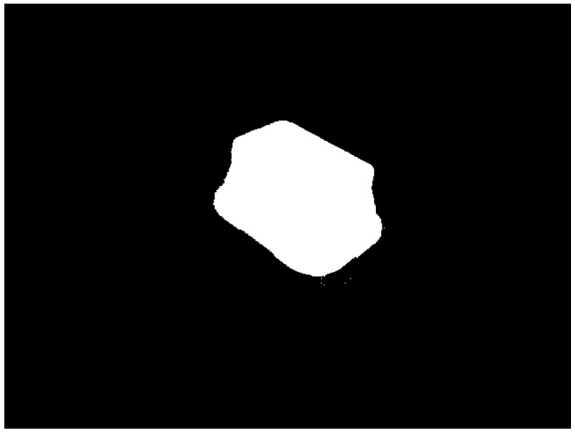
Image Source: *Hasenfratz, et al. (2004). "A Real-Time System for Full Body Interaction with Virtual Worlds"*

[1] *J. L. Schönberger and J. Frahm, "Structure-from-Motion Revisited"*
[2] *K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving,"*

Pipeline

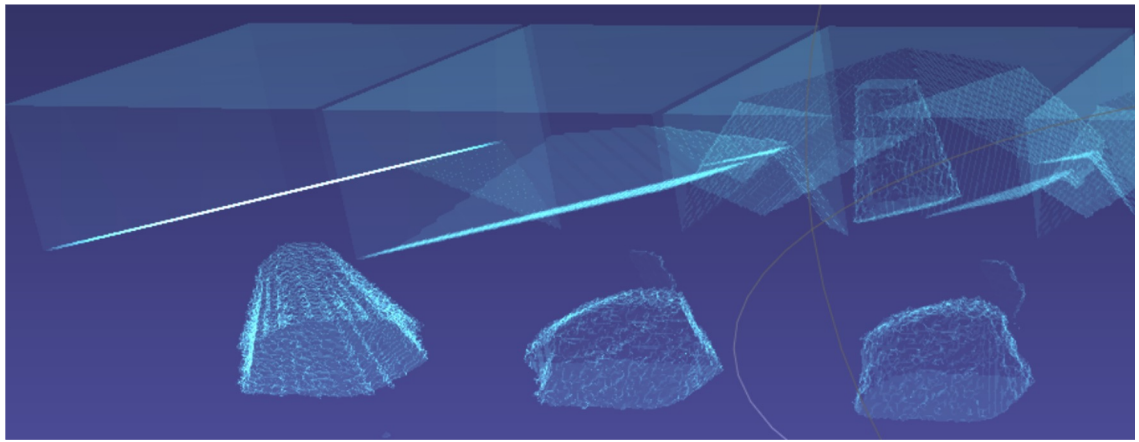
- Pose Estimation
- Image Segmentation
- Voxel Carving
- Post-Processing
- Marching Cubes





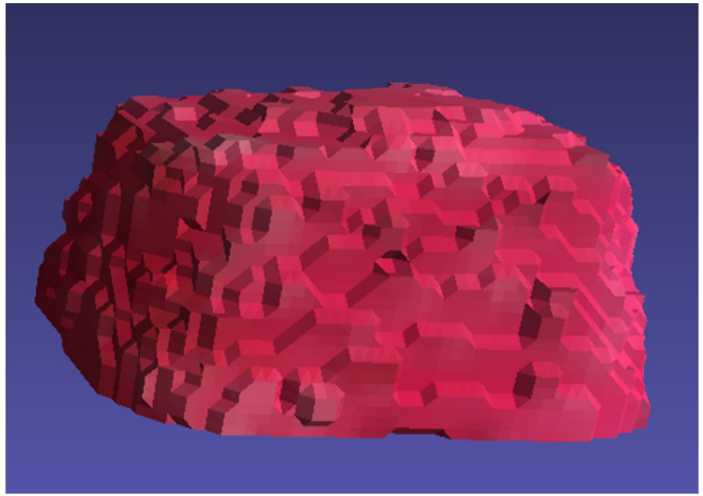
Pipeline

- Pose Estimation
- Image Segmentation
- **Voxel Carving**
- Post-Processing
- Marching Cubes



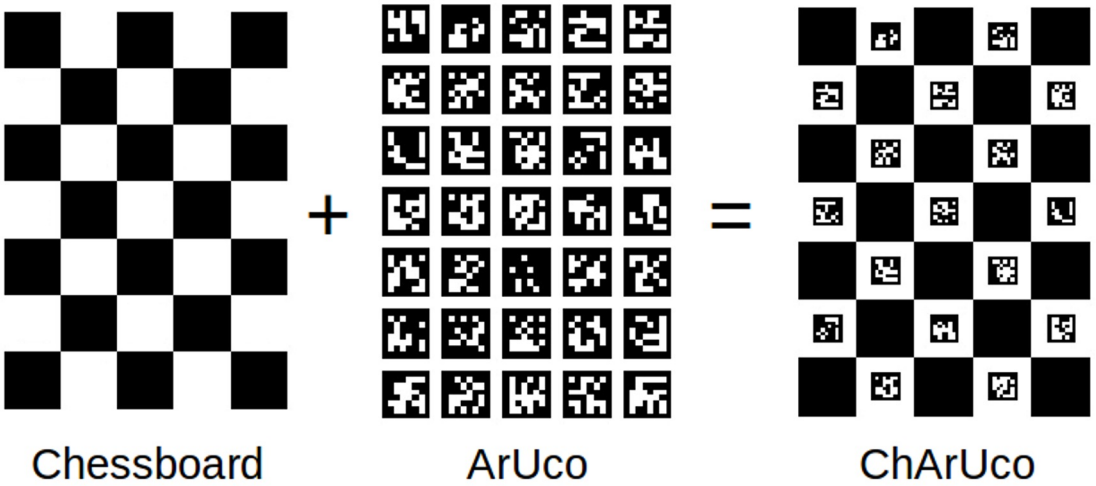
Pipeline

- Pose Estimation
- Image Segmentation
- Voxel Carving
- Post-Processing
- Marching Cubes



Voxel Carving Pipeline

Pose Estimation



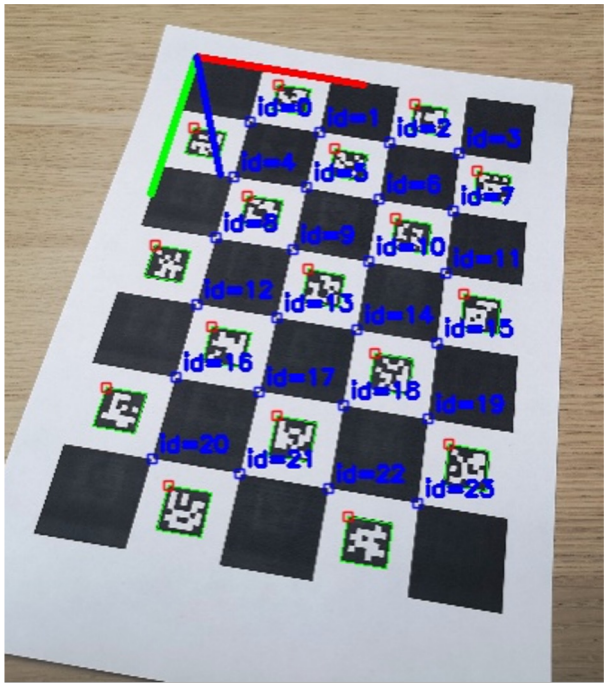
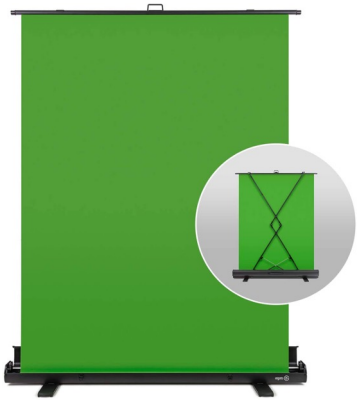
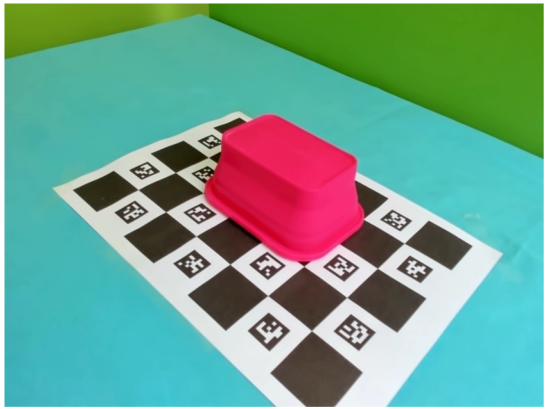


Image Segmentation





Source Image

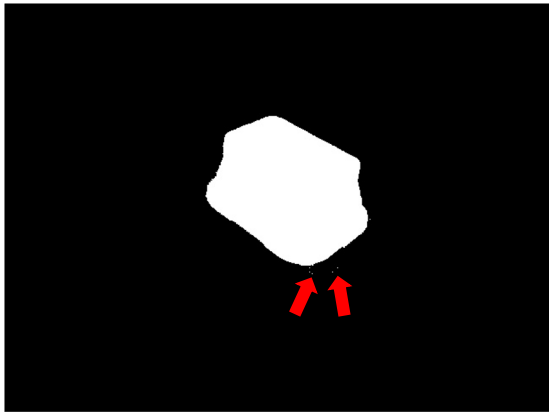


Image segmentation using RGB color masks

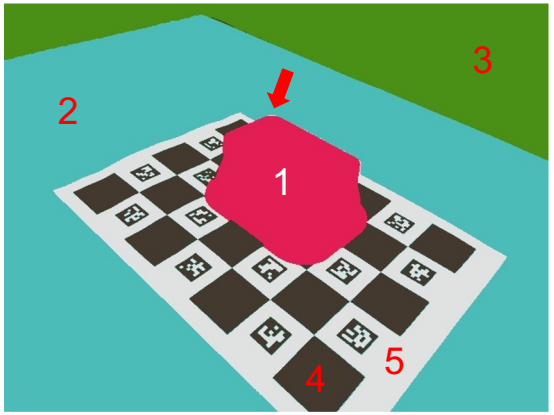
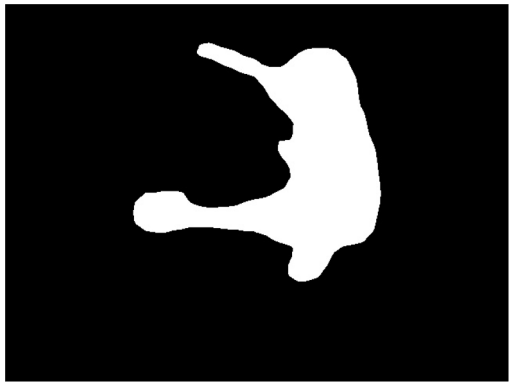


Image segmentation using K-means clustering

Image Segmentation

Source Image

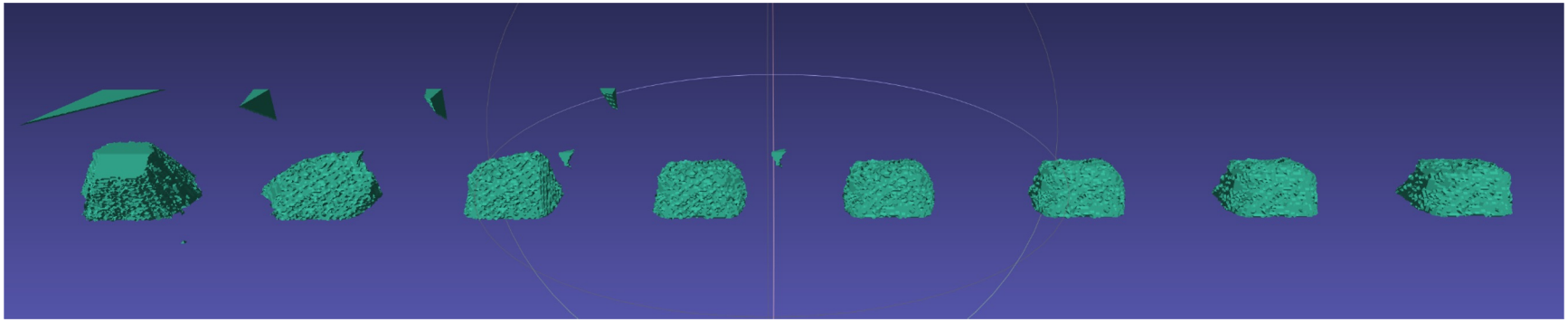


Segmentation using Mask R-CNN

Voxel Carving - Standard Approach

- User-defined voxel grid (e.g. dimensions 100x100x50, voxel size 0.0028m)
- Iteratively process images
- Projection: voxel (world-coord) to pixel (camera-coord) via pose estimation
- Segmentation → removal of voxels corresponding to pixels that are considered part of the background

⇒ *Θ(nxyz)* with *n : number of images, and x,y,z : dimensions of the voxel grid*



Voxel Carving - Greedy Approach

- processing queue (initially containing the origin of the voxel grid)
- for each voxel of queue: check if image exists, s.t. projection from world- to camera-coord maps to pixel that is determined to be part of the background.
  - no image exists → continue with next voxel
  - otherwise → remove voxel
- adjacent voxels revealed by the removal must be added to the processing queue

⇒ no processing of voxels encapsulated by the surface of the object

⇒ WC *O(nxyz)*

BUT: real world datasets result in significantly better performance

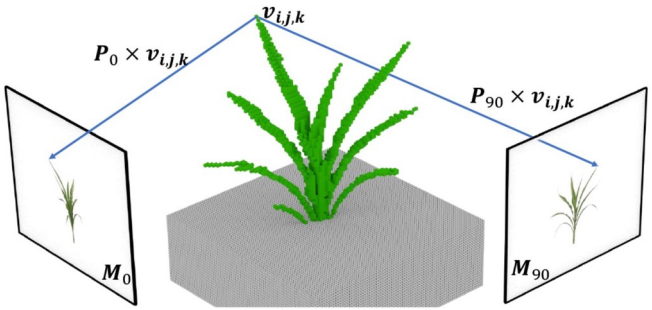
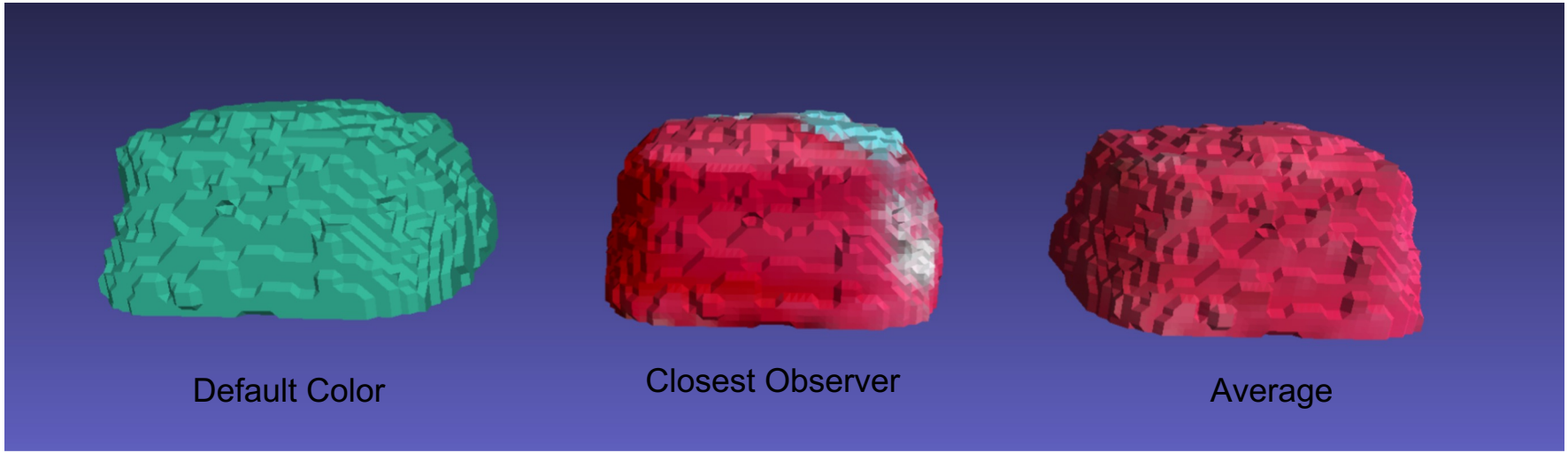


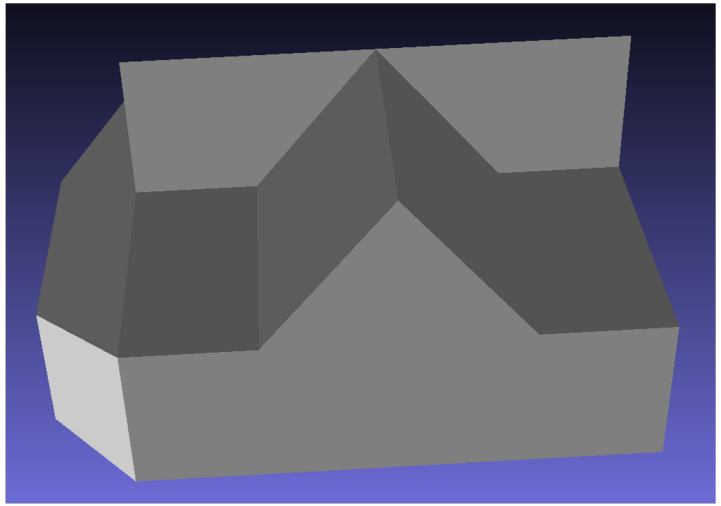
Image Source: *Gaillard, et al. (2020). "Voxel Carving Based 3D Reconstruction of Sorghum Identifies Genetic Determinants of Radiation Interception Efficiency"*

Color Reconstruction

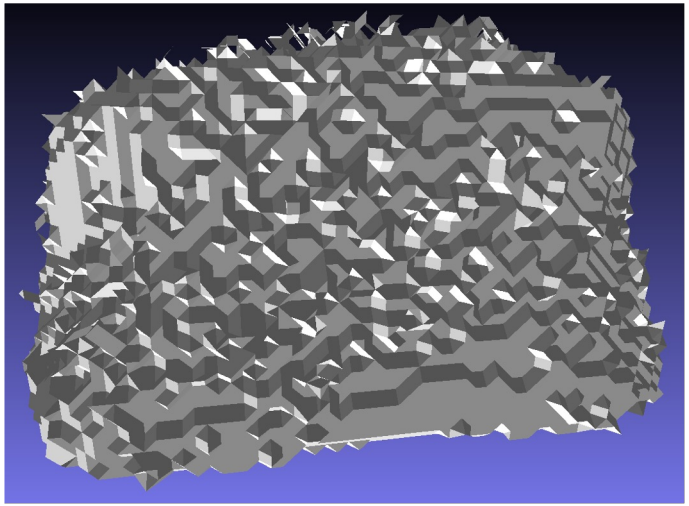


Color Reconstruction for the *box*-dataset

Postprocessing

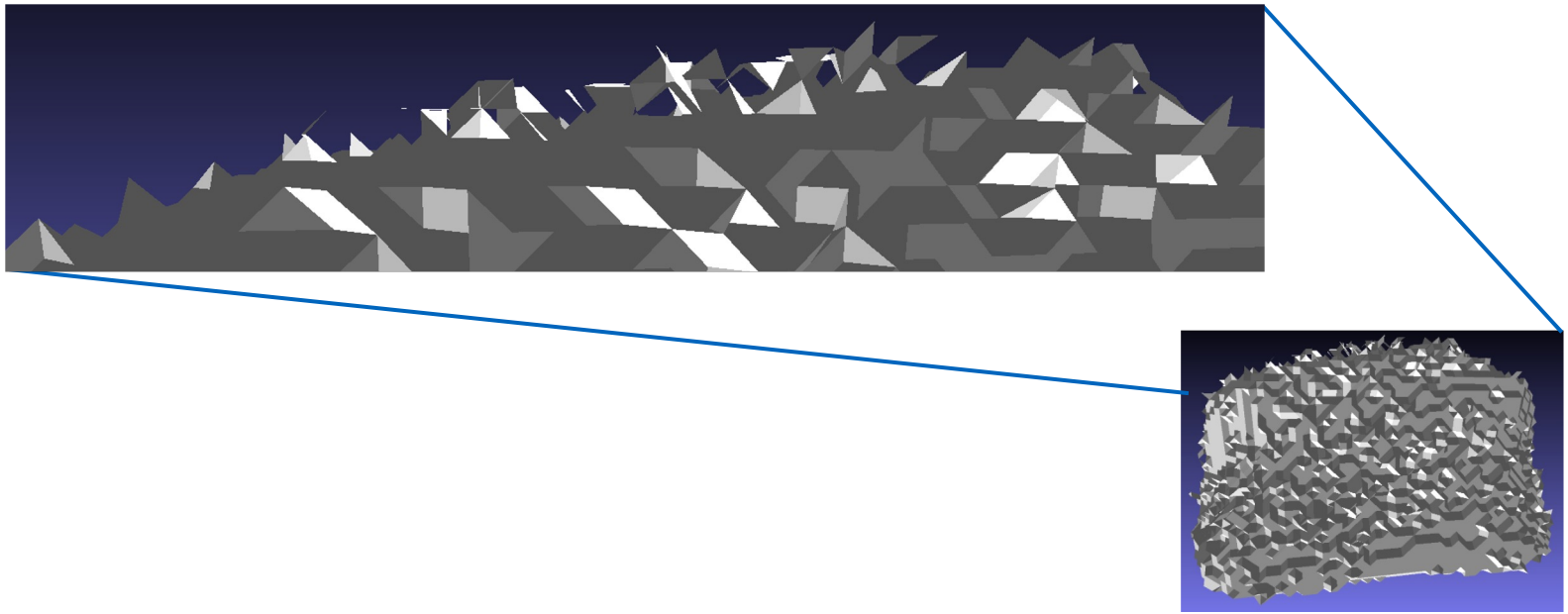


10x10x5 Voxels

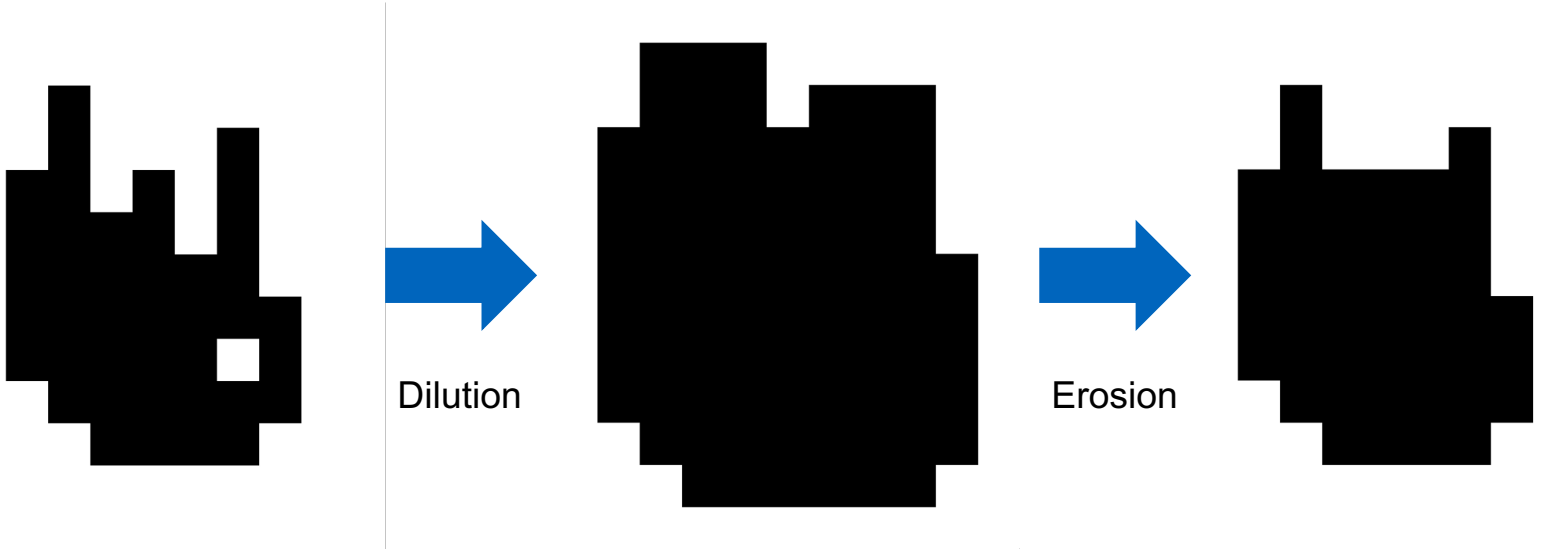


100x100x50 Voxels

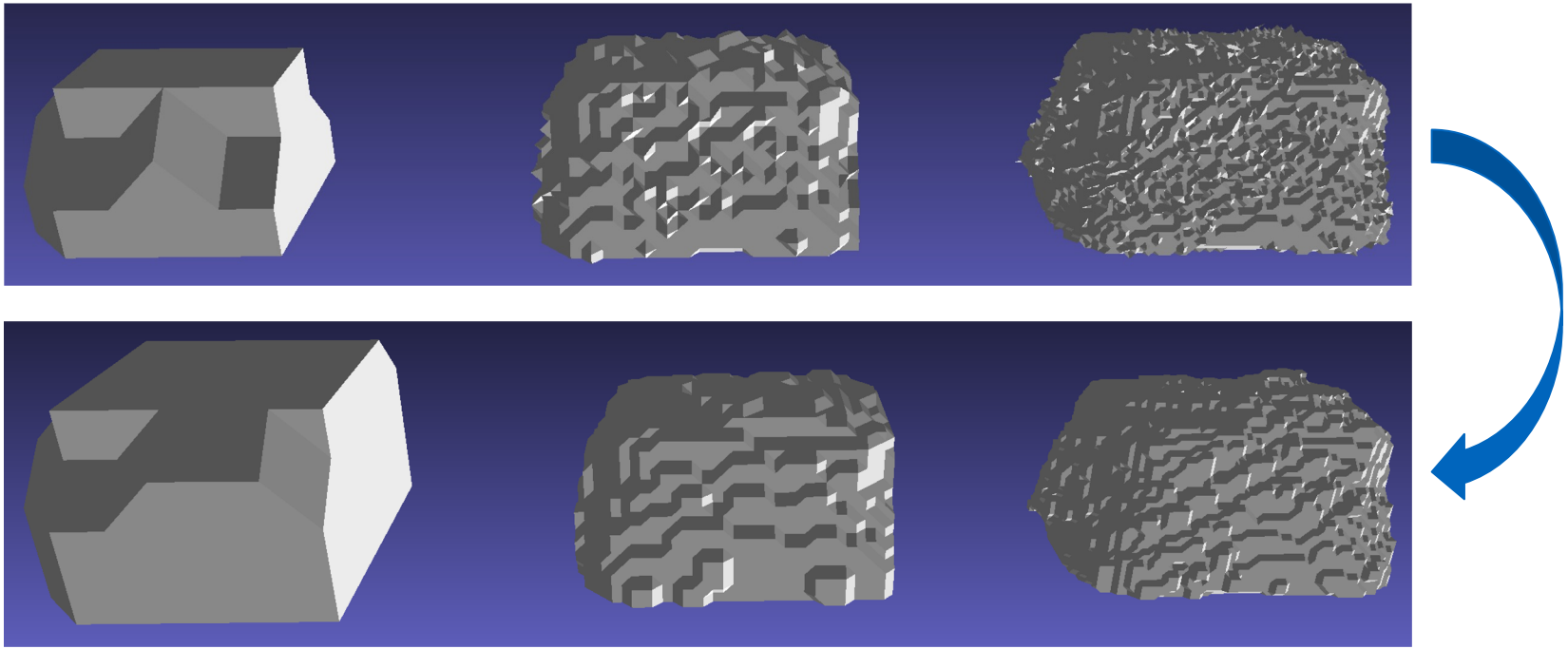
Postprocessing - Artifacts



Morphological Closing: Process

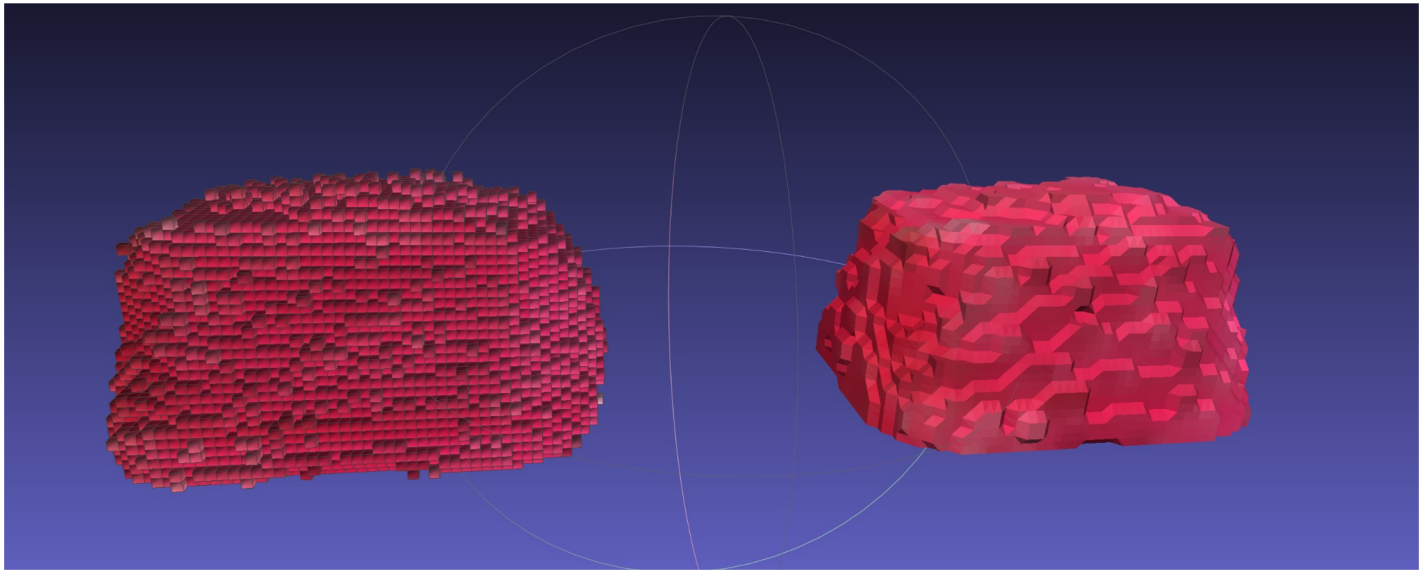


Morphological Closing: Result



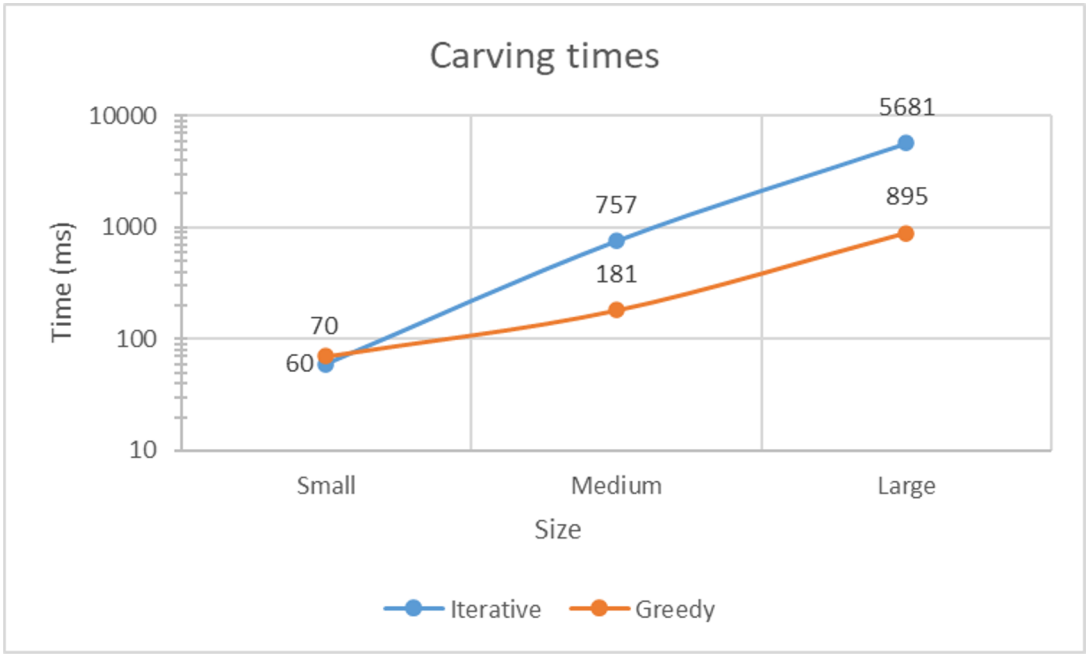
Model Conversion: Marching Cubes

internal voxel representation ➡ triangular surface mesh (.OFF-format)



Results

Performance

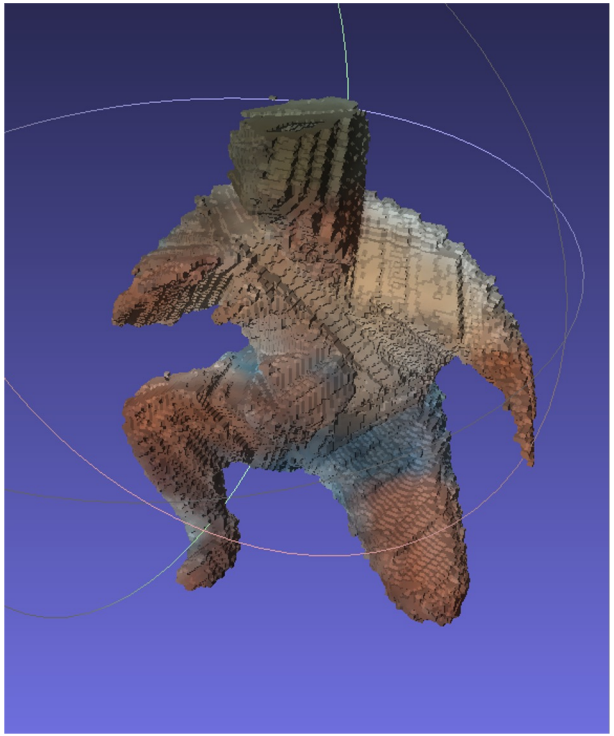


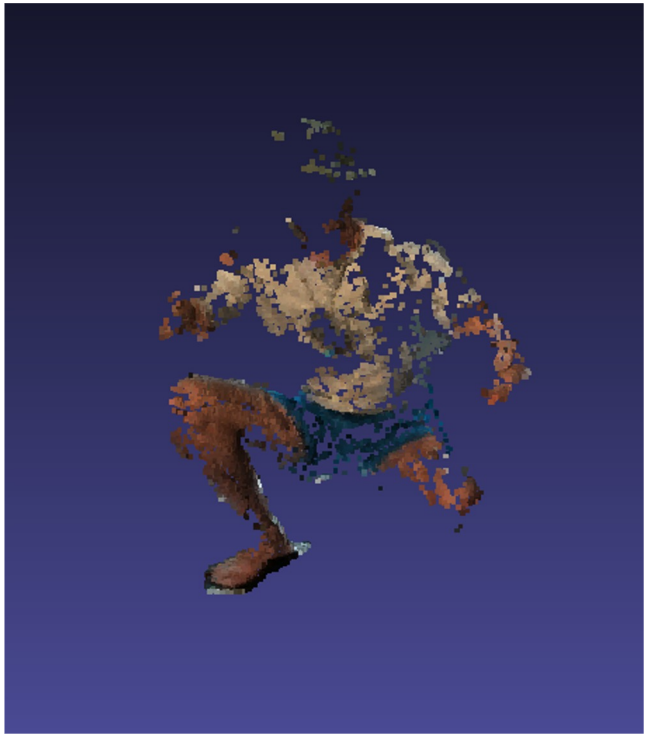
Box-data set

- n = 8
- small: 10x10x5
- medium: 50x50x25
- large: 100x100x50

- CPU: Intel Core i7-9700k @4.5GHz

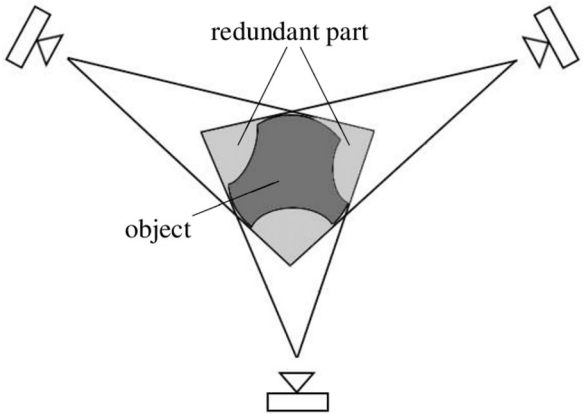
Human Model

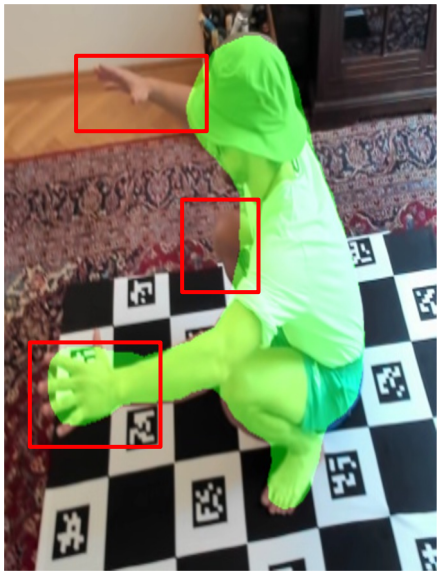




Our method vs. SfM (Colmap) on the human dataset (23 images)

Discussion





Future Work

Improve image segmentation

Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation

Feng Li[1,3*†], Hao Zhang[1,3*†], Huaizhe Xu[1,3], Shilong Liu[2,3], Lei Zhang[3‡], Lionel M. Ni[1,4], Heung-Yeung Shum[1,3]
[1]The Hong Kong University of Science and Technology.
[2]Dept. of CST., BNRist Center, Institute for AI, Tsinghua University.
[3]International Digital Economy Academy (IDEA).
[4]The Hong Kong University of Science and Technology (Guangzhou).
{hzhangcx,fliay,hxubr}@connect.ust.hk
{liusl20}@mails.tsinghua.edu.cn
{leizhang}@idea.edu.cn
{ni,hshum}@ust.hk

Abstract

In this paper we present Mask DINO, a unified object detection and segmentation framework. Mask DINO extends DINO (DETR with Improved Denoising Anchor Boxes) by adding a mask prediction branch which supports all image segmentation tasks (instance, panoptic, and semantic). It makes use of the query embeddings from DINO to dot-product a high-resolution pixel embedding map to predict a set of binary masks. Some key components in DINO are extended for segmentation through a shared architecture and training process. Mask DINO is simple, efficient, scalable, and benefits from joint large-scale detection and segmentation datasets. Our experiments show that Mask DINO significantly outperforms all existing specialized segmentation methods, both on a ResNet-50 backbone and a pre-trained model with SwinL backbone. Notably, Mask DINO establishes the best results to date on instance segmentation (54.5 AP on COCO), panoptic segmentation (59.4 PQ on COCO), and semantic segmentation (60.8 mIoU on ADE20K). Code will be avaliable at https://github.com/IDEACVR/MaskDINO.

1 Introduction

Object detection and image segmentation are fundamental tasks in computer vision. Both tasks are concerned with localizing objects of interest in an image but have different levels of focus. Object detection is to localize objects of interest and predict their bounding boxes and category labels, whereas image segmentation focuses on pixel-level grouping of different semantics. Moreover, image segmentation encompasses various tasks including instance segmentation, panoptic segmentation, and semantic segmentation with respect to different semantics, e.g. instance or category membership, foreground or background category.

Remarkable progress has been achieved by classical convolution-based algorithms developed for these tasks with specialized architectures, such as Faster RCNN [28] for object detection, Mask RCNN [12] for instance segmentation, and FCN [25] for semantic segmentation. Although these methods are conceptually simple and effective, they are tailored for specialized tasks and lack the

*Equal contribution.
†This work was done when Feng Li and Hao Zhang were interns at IDEA.
‡Corresponding author.

Preprint. Under review.

Statistical approach

A Probabilistic Framework for Space Carving

A. Broadhurst, T.W. Drummond and R. Cipolla
Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ

Abstract

This paper introduces a new probabilistic framework for ace Carving [9]. In this framework each voxel is as- ned a probability, which is computed by comparing the elihoods for the voxel existing and not existing.
This new framework avoids many of the difficulties asso- ated with the original Space Carving algorithm. Specifi- lly, it does not need a global threshold parameter, and it arantees that no holes will be carved in the model. This per also proposes that a voxel-based thick texture is a re- istic and efficient representation for scenes which contain minant planes. The algorithm is tested using both real d synthetic data, and both qualitative and quantitative ults are presented.

Introduction

Three-dimensional reconstruction has attracted consid- able attention, and is still a very active topic for re- arch. The goal of reconstruction is to recover the three- nensional shape of a static scene viewed from a number different cameras. Many different representations have en suggested, but still no single solution has been devel- ed that covers all types and configurations of input im- es.

The earliest attempts [10] to solve the reconstruction oblem posed the solution using image matching. These relation based algorithms (e.g. [10, 7, 15]) attempted solve the 2-D disparity field which described the rela- nship between two images. In this representation, when re is a change of depth, there is always an associated oc- sion. But, in all these algorithms the factor limiting the mplexity of the reconstruction is the depth parameterisa- n of space $z(x, y)$.

More recently, a second class of algorithms have focused using Euclidean (or projective) representations for space. common approach is to use piecewise planar polygons, hich can be made significantly more realistic by using photographic and bas-relief texture. This was demonstrated by [5], and further developed by [1] who used layered depth images to efficiently represent 3-D scenes. The algorithm in this paper uses a layer-based implementation.

A third class of algorithms are the volumetric represen- tations, which may use voxels [14, 13, 9, 4], or level sets [6]. The voxel-based algorithms make no attempt to model the continuity of shape, rather, they model the volume as an array of 3-D voxels. By not making assumptions about planarity and continuity, the voxel algorithms are able to cope with significantly more complex structures. The dif- ficulty with these approaches, however, is that by ignoring the regularising assumptions they become more susceptible to noise.

In this paper a novel probabilistic framework for the space carving algorithm will be presented. Many ideas from the original Space Carving [9] framework will be retained but, significantly the existence of a voxel will not be a bi- nary function. Instead each voxel will be assigned a proba- bility for it existing in the model. This is different from the work of [14, 2], because the probability is used to describe the existence of the voxel, and is not used to infer opacity. It will be shown that by using a probabilistic framework, two major failings of the Space Carving algorithm can be avoided. The first advantage is that no global parameters are required. (Space Carving requires the user to specify a global variance.) The second advantage, is that it can be guaranteed that the algorithm will not carve holes in the model, which is a well known failing of the Space Carving algorithm.

2. Background

In the Space Carving [9] framework, space is represented by an array of voxels, which must enclose the entire scene. At each iteration the algorithm selects a voxel and projects it into all the images where it is visible. It then asks the question: could this voxel have been part of the model? If the answer is no, then voxel is removed. This process is repeated until the algorithm has exposed the 3-D shape of

-7695-1143-0/01 $10.00 © 2001 IEEE

388

Real-time capable

Real-Time GPU-Based Voxel Carving with Systematic Occlusion Handling

Alexander Schick[1] and Rainer Stiefelhagen[1,2]

[1] Interactive Analysis and Diagnosis, Fraunhofer IITB Karlsruhe
alexander.schick@iitb.fraunhofer.de
[2] Institut für Anthropomatik, Universität Karlsruhe (TH)
stiefel@ira.uka.de,
rainer.stiefelhagen@iitb.fraunhofer.de

Abstract. We present an approach to compute the visual hulls of multiple people in real-time in the presence of occlusions. We prove that the resulting visual hulls are correct and minimal under occlusions. Our proposed algorithm runs completely on the GPU with framerates up to *50 fps* for multiple people using only one computer equipped with off-the-shelf hardware. We also compare runtimes for different graphic chips and show that our approach scales very well without additional effort. Comparison to other work shows that our algorithm is as fast as state-of-the-art technology. The resulting visual hulls can be the basis for a wide range of algorithms that require a robust voxel representation as input.

1 Introduction

We are interested in analyzing people that are interacting in *multi-camera environments* like smart rooms, smart control rooms, and smart housing. In these scenarios, the same scene is observed by multiple cameras. Due to different viewpoints, this leads to ambiguities. In one image, the actions of a person could be clearly visible, whereas in another image, the person might be occluded. Applications working with this data must find a way to reason with ambiguities, for example by fusing the inputs into one well-defined representation. Volumetric reconstruction algorithms, like voxel carving, provide a solution by building a coherent 3D representation.

The visual hull concept is well-known in the context of volumetric reconstruction and many applications can benefit from it. However, it relies on silhouette information and is very sensitive to segmentation errors. Static occlusions, like tables and chairs, cause severe problems for foreground segmentation and therefore for visual hulls in general. Current solutions do not address this problem properly. We will present a systematic solution to compute the minimal visual hulls during occlusions and prove its correctness.

To achieve real-time runtimes, many computer vision applications use the power of modern GPUs. By strictly following the highly parallel mathematical

J. Denzler, G. Notni, and H. Süße (Eds.): DAGM 2009, LNCS 5748, pp. 372–381, 2009.

Thank you for your attention!!!