Graphical Models 75 (2013) 126–136

Contents lists available at SciVerse ScienceDirect

Graphical Models

journal homepage: www.elsevier.com/locate/gmod

# Octree-based fusion for realtime 3D reconstruction

Ming Zeng, Fukai Zhao, Jiaxiang Zheng, Xinguo Liu *

*State Key Lab of CAD&CG, Zhejiang University, China*

ARTICLE INFO

*Article history:*
Received 29 August 2012
Received in revised form 18 September 2012
Accepted 25 September 2012
Available online 22 November 2012

*Keywords:*
Octree
KinectFusion
3D reconstruction
Graphics hardware
Signed distance function
Ray casting

ABSTRACT

This paper proposes an octree-based surface representation for KinectFusion, a realtime reconstruction technique of in-door scenes using a low-cost moving depth camera and a commodity graphics hardware. In KinectFusion, the scene is represented as a signed distance function (SDF) and stored as an uniform grid of voxels. Though the grid-based SDF is suitable for parallel computation in graphics hardware, most of the storage are wasted, because the geometry is very sparse in the scene volume. In order to reduce the memory cost and save the computation time, we represent the SDF in an octree, and developed several octree-based algorithms for reconstruction update and surface prediction that are suitable for parallel computation in graphics hardware. In the reconstruction update step, the octree nodes are adaptively split in breath-first order. To handle scenes with moving objects, the corresponding nodes are automatically detected and removed to avoid storage overflow. In the surface prediction step, an octree-based ray tracing method is adopted and parallelized for graphic hardware. To further reduce the computation time, the octree is organized into four layers, called top layer, branch layer, middle layer and data layer. The experiments showed that, the proposed method consumes only less than 10% memory of original KinectFusion method, and achieves faster performance. Consequently, it can reconstruct scenes with more than 10 times larger size than the original KinectFusion on the same hardware setup.



## 1. Introduction

Reconstruction of 3D scenes has been an important and active research area in computer vision and computer graphics. The traditional methods take advantages of 3D scanning hardware, depth camera, stereo vision and structured lights techniques to obtain the surfaces of the scene. Recently, Newcombe et al. [1] proposed a realtime reconstruction method for in-door scenes, called KinectFusion. They use Kinect, a low-cost moving depth camera, to capture the scene, and merge the captured depth maps into the final result. The scene is represented as a signed distance function (SDF) and stored in an uniform subdivided grid of voxels, so that they can use a commodity graphics hardware to parallelize the computation in tracking the pose of the Kinect and fuse the depth map in the scene, and achieve real time performance.

Since KinectFusion represents the signed distance function in an uniformly subdivided grid of voxels, it requires a lot of memory to store volumetric data. Meanwhile, most of the geometry of scene is essentially 2D manifold, and is very sparse in the scene volume. As shown in the experiments, most of the voxels in the signed distance function is empty, and are wasted consequently. When capturing larger scale scenes with finer details, the memory consumption will dramatically increases, which easily exceeds the memory limit of the commodity graphic hardware. For a resolution of $1024 \times 1024 \times 1024$, the grid data for the signed distance function takes 4 GB memory (assuming each voxel takes 4-byte for the distance value).

To address the above problem, we proposed an octree data structure to store the signed distance function of the

* Corresponding author.
*E-mail address:* xgliu@cad.zju.edu.cn (X. Liu).

1524-0703/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.gmod.2012.09.002

scene, and developed several algorithms to maintain the octree, and parallelize the computations in graphics hardware. The proposed algorithms greatly exploit the hierarchical structure of the octree and the GPU parallel computation ability. The proposed octree data structure allows for fast traversal, and consequently optimizes the reconstruction performance. In the step of reconstruction update, the octree is traversed in breadth-first order, while in the step of surface prediction, the octree is traversed in a top-down order, which can skip a lot of empty intermediate nodes. In experiments, the proposed method is much more efficient than the original KinectFusion in terms of memory consumption and running speed.

The main contribution of this paper is a memory-efficient and real-time scene scanning method, which consists of the following novel techniques:

- A memory-efficient and traversal-efficient octree data structure in GPU.
- Octree-based algorithms for reconstruction update and surface prediction.
- Several octree update algorithms for reconstructing dynamic scenes with moving objects.

## 2. Related work

### 2.1. Octree in GPU

There are several spatial data structures to organize 2D/3D data in computer graphics, e.g. kd-tree [2], BVH [3], and octree. Among these data structures, octree is used most widely, which appears in many applications, e.g. ray tracing [4], mesh simulation [5], mesh coding [6], and geometric modeling [7–9]. Octree adaptively splits the space of the scene according to the distribution of geometric primitives of the scene, and drops the empty cells to save the storage cost. As the development of programmable graphics hardware, octree can be built and traversed completely in GPU, and take advantage of the parallel computing feature of GPU to dramatically improve the performance of rendering and geometry processing. Compared with the traditional methods, we deal with time varying octrees.

### 2.2. 3D reconstruction

There have been many kinds of systems for capturing 3D scenes based on stereo vision, structured lights, 3D laser scanner and depth camera. Stereo vision based methods infer the camera parameters from a set of images of the target scene, match the feature points across the images, and then compute the 3D positions of the matched feature points, recovering a sparse point cloud of the scene [10,11]. A survey and comparison of stereo methods can be found in [12]. Stereo algorithms have been successfully applied to reconstruct large scale scenes and small scale objects, such as buildings, human body, and faces Stereo, but it is still challenging to handle varying lighting, non-diffuse surface. And the stereo methods are timing consuming and not suitable for real time applications. People in Robotic community focus on simultaneous localisation and mapping (SLAM) [13,14], which can track the camera in real time and generate a sparse point cloud of the scene. In order to obtain dense reconstruction result in real time, people combined SLAM with multiview stereo techniques [15,16]. Recently, Newcombe et al. used iterative image alignment method to obtain the correspondence, instead of using the traditional image feature points, which can densely reconstruct a small scene (about $1\text{ m} \times 1\text{ m} \times 1\text{ m}$) in real time [17].

To meet the requirement of real time applications, depth cameras based on time-of-flight and structured light are used to capture a set of depth maps of the scene from different view points, which are then aligned together and fused into a complete surface representation. To align the depth maps, there are a number of registration methods [18,19] based on the iterative closest point (ICP) method proposed by Chen and Medioni [20]. ICP method is originally developed for rigid objects, and can be extended to handle deformable objects [21–24].

Recently, Newcombe et al. [1] developed a real time reconstruction method for in-door scenes, called KinectFusion, using a low-cost moving depth camera and commodity graphics hardware. The basic idea of KinectFusion is to maintain a volumetric scene data to align and merge the depth map streamed from the Kinect. KinectFusion can also be used for camera tracking and user interaction in augmented reality [25]. In order to handle unbounded scenes, Whelan et al. [26] extended KinectFusion to "Kintinuous" by shifting the volume and extracting meshes continuously.

Zeng et al. [27] introduced an octree structure for KinectFusion to save the memory cost of KinectFusion and achieve faster performance. Compared with "Kintinuous" [26], this method improves the core parts of the KinectFusion, which is memory-efficient and can be a building block of large scale scanning systems in "Kintinuous". In this paper, we extend the work of Zeng et al. [27] by (1) introducing node removal operation to handle dynamic scenes with moving objects; (2) adding a top layer to the octree for accelerating the traversal in the surface prediction step; and (3) presenting detailed experimental results and analysis on the proposed method.

## 3. Overivew

KinectFusion captures the depth map of the scene using Kinect, a low-cost depth camera, and successively fuses the depth map into the reconstructed scene. It represents the scene as the zero valued surface of a signed distance function (SDF) and stores the SDF in an uniformly subdivided grid of voxels. It tracks the pose of the camera, so that the depth map can be correctly positioned in the volumetric grid.

As shown in Fig. 1, KinectFusion consists of four main steps:

- **Surface Measurement** converts the depth map streamed from the depth camera into a point cloud and estimates the normal of each point.