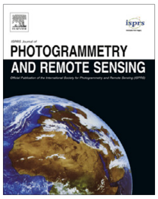
ARTICLE IN PRESS

ISPRS Journal of Photogrammetry and Remote Sensing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



Geometrically stable tracking for depth images based 3D reconstruction on mobile devices

Yangdong Liu, Wei Gao*, Zhanyi Hu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 1 October 2017
Received in revised form 31 January 2018
Accepted 8 March 2018
Available online xxxx

Keywords:
Visual-SLAM
Dense mapping
ICP algorithm
Sampling algorithm
On-line
Mobile device

ABSTRACT

With the development of hardwares such as mobile devices and portable depth cameras, on-line 3D reconstruction on the mobile devices with depth streams as input turns to be possible and promising. Most existing systems use volumetric representation methods to fuse the depth images and use ICP algorithm to estimate the poses of cameras. However, ICP tracker suffers from large drift in scenes containing insufficient geometric information. To deal with this problem, we propose a stability based sampling method which select different number of point-pairs in different windows according to their geometric stability. In addition, we fuse the ICP tracker with the IMU information through an analysis of the condition number. Then we apply the stability based sampling method to the spatially hashed volumetric representation. Qualitative and quantitative evaluations of tracking accuracy and 3D reconstruction results show that our method outperforms the current state-of-the-art systems, especially in scenes lacking sufficient geometric information. In total, our method achieves frame rates 20 Hz on an Apple iPad Air 2 and 200 Hz on a Nvidia GeForce GTX 1060 GPU.

1. Introduction

Visual simultaneous localization and mapping (V-SLAM) aims to simultaneously estimate the motion of a robot or a camera and to reconstruct the geometric structure of the unknown environment that the device is observing. It is proposed in the robot domain initially and then used as a key technique in both the robotics and augmented reality (AR). After years of research, some V-SLAM techniques such as feature-based V-SLAM tend to be mature (Mur-Artal et al., 2015; Klein and Murray, 2007). While these methods only construct a sparse representation of the environment, high-resolution dense mapping needs to be further studied. Dense mapping aims to obtain a high quality 3D reconstruction of the scenes. With the popularity of mobile devices such as mobile phones and tablet computers, estimation of the motion and 3D reconstruction of the environment using a mobile device turn to be promising in the future. Dense on-line reconstruction of indoor scenes on mobile devices opens up many useful applications including 3D scanning of interesting objects and AR. However, mobile devices have limited computational resources, which makes on-line and realistic 3D reconstruction on mobile devices remain an unsolved problem. For this reason, this work aims at on-line 3D reconstruction on mobile devices.

Many recent systems have been put forward to obtain real-time 3D reconstruction. 3D reconstruction with monocular cameras needs to recover depth information passively, which suffers from high computational complexity. Besides, the calculated depth images are quite noisy especially in regions with weak color texture. The consuming-level depth cameras as the Microsoft Kinect (Microsoft, 2010) enable consumers to obtain depth information actively. KinectFusion (Newcombe et al., 2011; Izadi et al., 2011) fuses depth images into the volumetric representations (Curless and Levoy, 1996), which proves to be a powerful method for generating dense, realistic 3D models. Thanks to its computational efficiency and algorithmic simplicity, the volumetric representation method has been reimplemented lots of times and leads to a wide range of further researches. Recently, devices such as Google Tango (Google, 2014) and Occipital Structure Sensor (Occipital, 2014) make 3D reconstruction on mobile devices practicable. Due to the convenience of capturing depth images and the concerns on geometric structure of the environment, we use the depth

* Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: wgao@nlpr.ia.ac.cn (W. Gao).

https://doi.org/10.1016/j.isprsjprs.2018.03.009
0924-2716/© 2018 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

Please cite this article in press as: Liu, Y., et al. Geometrically stable tracking for depth images based 3D reconstruction on mobile devices. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.009

streams as the input to reconstruct geometric models of indoor scenes in this work.

In this paper, we introduce a geometrically stable tracking and dense mapping method which can perform 3D scanning at a speed of up to 20 Hz on an Apple iPad Air 2 and 200 Hz on a Nvidia GeForce GTX 1060 GPU. One of our contributions is to apply a stability based sampling method to the iterative closest point (ICP) algorithm (Besl et al., 1992; Blais and Levine, 1995; Chen and Medioni, 1991). Through extensive experiments we demonstrate that our method is of higher accuracy, especially in scenes with less geometric features compared to the current state-of-the-art methods. Our second contribution lies in the fusion of ICP tracker with inertial measurement unit (IMU) information through an analysis of the condition number. The third contribution is that we build uncertainty maps of the depth images captured by an Occipital Structure Sensor and use them to adaptively determine truncation distances during volumetric integration.

1.1. Related works

With the popularity of inexpensive depth cameras, Newcombe et al. propose a dense 3D reconstruction framework named KinectFusion (Newcombe et al., 2011; Izadi et al., 2011), which opens up an era of high-quality and real-time 3D reconstruction. KinectFusion makes use of a volumetric data (Curless and Levoy, 1996) to represent the scenes. A truncated signed distance function (TSDF) value and its weight are stored in every voxel of the volume. The TSDF value is the distance between the center of a voxel and the nearest surface of the observed object. Camera poses are determined by a frame-to-model ICP method, which is followed by a simple weighted running average (Curless and Levoy, 1996) to fuse the incoming depth images into the volumetric data. The surface is encoded into an implicit function and the surface mesh is extracted by marching cubes (MC) algorithm (Lorensen and Cline, 1987).

Though KinectFusion has many advantages such as its computational efficiency and algorithmic simplicity, it has some disadvantages of the representation method and the tracking method. For the representation method: the 3D reconstruction lacks scalability because the volume is predefined; the occupied memory increases with the entire space rather than with the surface area; the voxels are uniformly divided, which cannot satisfy the multi-resolution representation of the scenes. For the tracking method: the effectiveness of ICP algorithm depends on the richness of geometric features; due to error accumulation the loop cannot be closed.

Various methods have been put forward to overcome these disadvantages. Many researchers propose the moving volume method (Roth and Vona, 2012; Whelan et al., 2015a). Voxels in the field of view are stored and processed in the device memory, while other voxels are turned into meshes and transferred to the long-term memory. This procedure is irreversible and lossy. Other methods aim to allocate and update the voxels around the actual surface (Zeng et al., 2013; Chen et al., 2013; Steinbrücker et al., 2014; Nießner et al., 2013; Kähler et al., 2015, 2016b). Octrees (Zeng et al., 2013; Chen et al., 2013; Steinbrücker et al., 2014) and hash tables (Nießner et al., 2013; Kähler et al., 2015, 2016b) are applied to retrieve the allocated voxels. These methods reduce computational cost and spare memory occupation. In order to represent the scene at multi-resolution, Henry et al. propose a patch volumes method which divides the entire scene into several volume blocks of various size and resolution (Henry et al., 2013). These volumes are aligned with dominant planes. Kähler et al. propose a hierarchical voxel block hashing method (Kähler et al., 2016a). This method is able to represent the observed space using higher resolution for parts which require more detailed representation. In consideration of limited computational resources, we use the voxel hashing method to retrieve the allocated voxels in our system.

In order to reduce accumulated error, some researchers try to utilize other kinds of sensors such as RGB cameras and IMUs. Fovis is based on the sparse features in the RGB images and the tracking result is affected by the abundance of color-texture (Huang et al., 2011). Dense Visual Odometry (DVO) uses depth information and RGB information simultaneously to obtain camera pose (Steinbrücker et al., 2011), which is appropriate for those cases with small relative poses. Whelan fuses the Fovis, DVO and ICP tracking methods altogether to lessen the dependence on color-textures and geometric features (Whelan et al., 2013). Nießner et al. use the integration of angular velocity as the initialization of rotational components in ICP algorithm (Nießner et al., 2014). Tanskanen et al. fuse the IMU measurements and vision measurements with Extended Kalman Filter (EKF) (Tanskanen et al., 2015). Li and Mourikis propose MSCKF to fuse multi-sensor measurements (Li and Mourikis, 2013). In our system we use the integration of angular velocity as the initialization of rotational components for ICP.

Based on the above methods, some systems aim to achieve the goal of 3D reconstruction on mobile devices using RGB-D images and IMU information. Kähler et al. introduce a system named InfiniTAM which integrates depth images at very high frame rates (Kähler et al., 2015, 2016b). They make substantial improvements on the voxel hashing method (Nießner et al., 2013) and achieve a rate of up to 20 Hz when processing IMU augmented 320 × 240 depth images on Apple iPad Air 2. However, InfiniTAM suffers from large drift if the depth images contain insufficient geometric features. Klingensmith et al. propose CHISEL which enables house-scale dense 3D reconstruction on a Google Tango (Klingensmith et al., 2015). They use the voxel hashing to represent the scene and combine visual-inertial odometry (VIO) with ICP to track the camera.

1.2. System outline

In accordance with what is widely used in previous works, we use the volumetric representation method to integrate depth images. Our system is composed of four main units as what is proposed in KinectFusion, which is shown in Fig. 2.

(a) Preprocessing: When a depth image is input, a dense vertex map and a normal map in the camera coordinate system are generated. Additionally, an uncertainty map representing the standard deviation (STD) of the depth noise and a gradient map used to determine the depth discontinuities are calculated as well.

(b) Camera Tracking: We register the input depth image and the ray-casted depth image from the proceeding camera pose through the well-known ICP algorithm to get a 6-DoF rigid relative transformation between them. If IMU is available, we fuse ICP tracker with IMU information.

(c) Volumetric Integration: After estimating the camera's global pose, depth images are integrated into a TSDF model through running average. The truncation distance is adaptive according to the noise level of measured depth.

(d) Surface Prediction: Ray-casting the volumetric model into a predicted surface is the final unit. This predicted surface is aligned with the live depth image in the tracking stage and provided to the user for visualization.

Each of the above units is elaborate in the following sections. One example of our reconstruction result on the real-world scene is shown in Fig. 1

Please cite this article in press as: Liu, Y., et al. Geometrically stable tracking for depth images based 3D reconstruction on mobile devices. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.009