Computer Vision and Image Understanding 000 (2016) 1–13

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Large scale and long standing simultaneous reconstruction and segmentation

Keisuke Tateno[a,b,*], Federico Tombari[a,c], Nassir Navab[a,d]

[a] Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Munich Germany
[b] Canon Inc., Shimomaruko, Tokyo Japan
[c] Dipartimento di Informatica: Scienza e Ingegneria (DISI), University of Bologna, Bologna Italy
[d] Johns Hopkins University, Baltimore USA

ARTICLE INFO

Article history:
Received 7 December 2015
Revised 8 May 2016
Accepted 24 May 2016
Available online xxx

Keywords:
Dense SLAM
Segmentation
Real-time
Scalable
Long standing
Relocalization
Loop-closure

ABSTRACT

This work proposes a method to segment a 3D point cloud of a scene while simultaneously reconstructing it via Simultaneous Localization And Mapping (SLAM). The proposed method incrementally merges segments obtained from each input depth image in an unified global model leveraging the camera pose estimated via SLAM. Differently from other approaches, our method is able to yield segmentation of scenes reconstructed from multiple views in real-time and with a complexity that does not depend on the size of the global model. Moreover, we endow our system with two additional contributions: a loop closure approach and a failure recovery and re-localization approach, both specifically designed so to enforce global consistency between merged segments, thus making our system suitable for large scale and long standing reconstruction and segmentation. We validate our proposal against the state of the art in terms of computational efficiency and accuracy on several benchmark datasets, as well as by showing how our method enables real-time reconstruction and segmentation of diverse real indoor environments.

1. Introduction and related work

Scene segmentation is one of the most important and researched topics in the field of robotic perception, since segmentation is typically a pre-requisite for several robotic tasks such as object modeling and object recognition (Aldoma et al., 2013), autonomous grasping and manipulation of objects (Kenney et al., 2009), object tracking (Mörwald et al., 2009), and scene understanding and object discovery of unknown environments (Karpathy et al., 2013). Within the robotic perception and computer vision communities, a great effort has been made to develop efficient 3D segmentation algorithms, i.e. real-time processing of depth maps obtained from RGB-D or 3D sensors. The focus on 3D data is motivated by the additional insight that geometry and shape provide, with respect to texture and color, for the task of segmentation, as well as the opportunity to determine segments that lie in the 3D space in front of the robot and not just on the image plane. Fast real-time segmentation of depth maps has been recently investigated by the works of Uckermann et al. (2012, 2013), Pieropan and Kjellstrom (2014) and Abramov et al. (2012).

Recently, 3D reconstruction methods, which aim at a real-time registration of depth maps from multiple viewpoints obtained from a moving sensor, are becoming increasingly exploited for higher level robotic perception tasks, since they offer additional information for the surrounding environment and are fundamental for robot navigation tasks: this is the case of Kinect Fusion (Newcombe et al., 2011), as well as dense SLAM (Henry et al., 2012; Kerl et al., 2013; Whelan et al., 2014). While the former method yields a 3D mesh of the reconstructed environment by exploiting a specific data representation internally deployed, the output of SLAM methods is generally in the form of a 3D point cloud.

As a consequence, in addition to segmentation methods aimed at processing single depth maps, some works have recently addressed the problem of segmenting 3D reconstructions obtained via Kinect Fusion or SLAM. Toward this goal, segmentation methods specifically devised to work on 3D meshes (Felzenszwalb and Huttenlocher, 2004) or point clouds (Golovinskiy and Funkhouser, 2009; Rabbani et al., 2006; Strom et al., 2010) are generally deployed to yield a segmentation of such 3D representations, as proposed in the object discovery approach of Karpathy et al. (2013). One main limitation of such methods is clearly the computational cost, since they cannot run in real-time. This aspect strongly limits their use in those application scenarios characterized by real-time constraints, as it is the case in most of the aforementioned robotic

* Corresponding author.
E-mail address: tateno@in.tum.de (K. Tateno).

http://dx.doi.org/10.1016/j.cviu.2016.05.013
1077-3142/

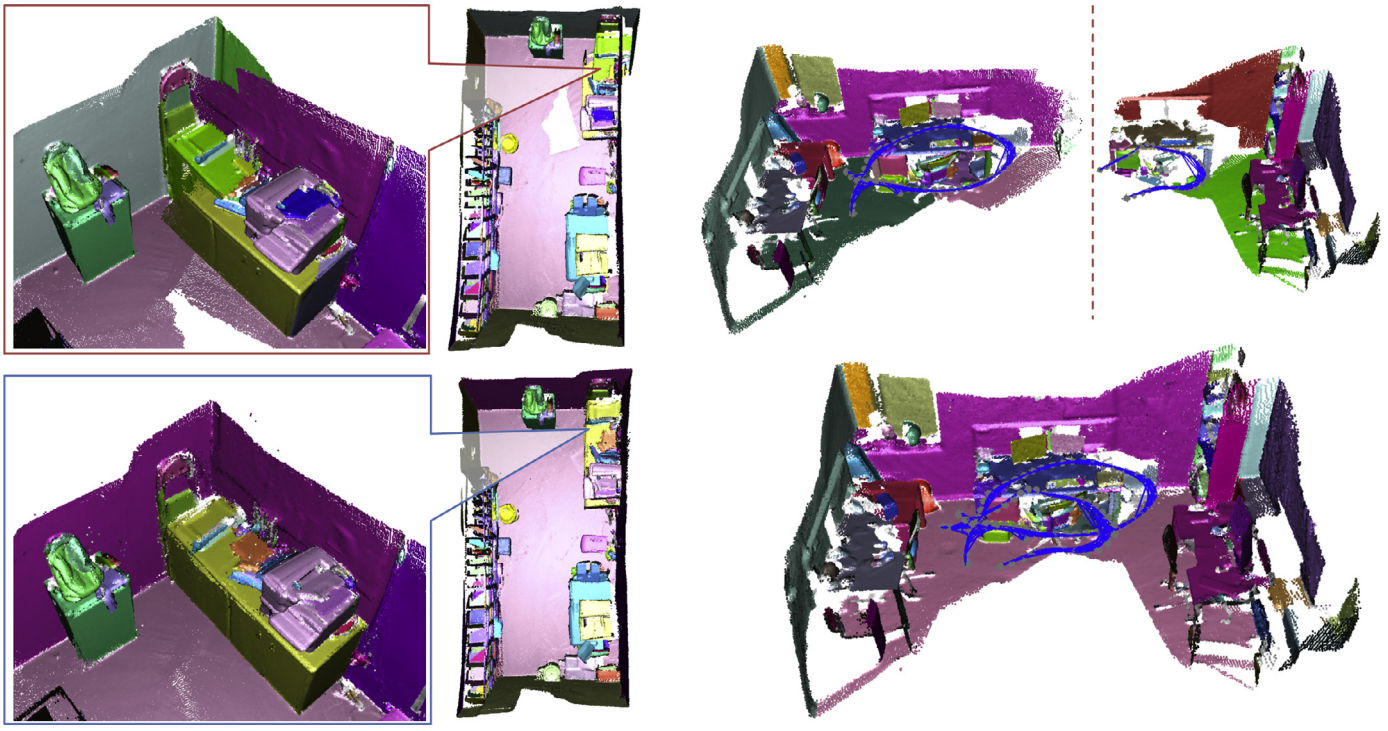

Fig. 1. The proposed framework is capable of providing a long-standing SLAM reconstruction embedding a globally consistent segmentation of the scene (bottom, left and right). It involves specific approaches for loop closure and failure recovery, which can enforce global label consistency among the segments while, respectively, globally optimizing the camera poses and handling tracking failures. This can be seen by comparing the top left image (before loop closure) with the bottom left one (after loop closure). In addition, an example of our failure recovery is portrayed in the bottom right image, which shows the merged result of the two independent sequences shown in the top right image. Blue lines depicts the edges of the key-frame graph, while grey points represent the estimated camera poses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

perception tasks. In addition, their computational burden tends to increase with the size of the 3D mesh or point cloud. Hence, to limit the overall computational requirements, only up to a certain number of merged depth maps can be deployed with an off-the-shelf hardware.

Aiming at the same goal, Finman et al. (2014) proposes incremental object segmentation in a dense RGB-D SLAM framework. The method is based on Kintineous (Whelan et al., 2014), a dense RGB-D SLAM approach which builds upon KinectFusion (Newcombe et al., 2011), and relies on a 3D representation called Truncated Sign Distance Function (TSDF). In this method, newly merged depth frames in the TSDF are, from time to time, extracted in the form of "slices" (i.e., a 3D mesh) according to the estimated camera position, segmented via graph-based segmentation (Felzenszwalb and Huttenlocher, 2004), and merged into a global segmentation map. Although this yields a much higher efficiency than the previous approaches, the use of a segmentation method such as (Felzenszwalb and Huttenlocher, 2004) on each "slice" (represented as a 3D mesh), as well as the fact that the segments extracted from each slice are successively merged in the global segmentation map, still does not allow this method to produce the segmentation of the current input data in real-time, and, as stated in Finman et al. (2014), generates an overall computational complexity that grows with the size of the global segmentation map.

Moreover, the work by Salas-Moreno et al. (2014) aims at real-time plane segmentation and SLAM reconstruction. In this method, planes are segmented from each input depth map, then they are incrementally merged into a global model. The main limitation of such a method in the context of segmentation is the fact that it considers only planar surfaces, while curved surfaces are not segmented, posing a limit to the generality of the processed shapes especially in the presence of arbitrary objects.

To tackle this problem, Finman and J. J. Leonard (2015) proposed a SLAM-based place recognition method based on object detection. The method recognizes the same place on different maps by detecting object segments that are common between partially overlapping maps. Although this method can attain label consistency of the objects appearing in different maps, it is still limited to specific shapes, since the object detection method relies on the objectness which cannot detect relatively flat surfaces. Moreover, the segment merging procedure is only partial, since it is carried out only on those segments which are recognized as objects.

Our approach aims at overcoming the limitations of the methods currently available in literature by simultaneously performing reconstruction and segmentation. In particular, while reconstruction is carried out within a point-based fusion SLAM framework so to deal with noisy 3D data, segments are extracted from the current depth frame and incrementally merged together so to build a Global Segmentation Map (GSM) of the reconstructed environment. As a consequence, it achieves the important advantage of a constant runtime regardless of the size of the GSM and the number of merged depth maps in the global 3D model, which makes our approach particularly suited to large-scale and long-standing reconstruction scenarios. Furthermore, at each frame, the update procedure is efficient enough to be carried out in real-time.

In addition to this approach, our work also includes two additional contributions, aimed at making our system apt to deal with large-scale and long-standing acquisitions. The first is a peculiar loop closure stage, that, when loop closures are detected, aims at obtaining a globally refined alignment of the camera poses and of the segment labels so to yield a reconstructed point cloud which is globally consistent not only in terms of 3D geometry, but also in terms of 3D segments. An example is shown in Fig. 1, which illustrates the difference before (top image, left) and after (bottom image, left) applying the proposed label-consistent loop closure

Please cite this article as: K. Tateno et al., Large scale and long standing simultaneous reconstruction and segmentation, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.05.013