SemanticPaint

Adam Kosiorek *

Abstract. The short abstract (50-80 words) is intended to give the reader an overview of the work.

1 Introduction

Capturing your own environment has never been easier. SemanticPaint can register your surroundings which, after undergoing a low-level 3D reconstruction, can be semantically segmented in an interactive way. Not only it works in real time but also requires no pretraining. Adding new object categories on the fly is facilitated by online model updates. The user is provided with instantaneous feedback and can re-label any object to correct errors. SemanticPaint makes capturing customized environment models with object classes particular to the user's interest easy and efficient.

The pipeline starts with capturing the environment as a stream of noisy RGBD images and combining them into a updated 3D model in an online fashion. The user can choose which objects to label and can do so by "touching" a small part of an object or encircling one with his hand and uttering the label. It is recognized by a standard speech recognition system. The label and the information about the affected data points are further passed to a Streaming Random Forest classifier which constantly learns and labels all visible voxels. To further improve classification results a spatially dense labeling is produced by an efficient mean-field inference algorithm. One of the biggest strength of SemanticPaint is the efficiency of each part of the pipeline, which translates to real time performance. Algorithms used in the pipeline were adapted to work on volumetric data in the TSDF format directly in order to avoiding the costly conversions to mesh or point-cloud formats. To allow this, the Voxel Oriented Patch features — a new type of a discriminative feature describing the voxel space — has been designed. The contributions can be summarized as follows: 3D semantic modeling system, Streaming Random Forest, efficient Mean-Field inference, Voxel Oriented Patch features.

Numerous applications are possible: (1) building large scale datasets of 3D objects or whole scenes for use in large-scale computer vision systems (2) using the dense semantic labeling of 3D environments in robot navigation or to aid people with impaired sight and (3) map environments for use in augmented reality scenarios or games.

The rest of the paper is organized as follows: Section 2. describes the related work, section 3 details internal data handling, section 4 describes the efficient

* Advisor: M.Eng. Keisuke Tateno, Chair for Computer Aided Medical Procedures & Augmented Reality, TUM, WS 2015/16.

mean-field inference algorithm, section 5 outline the Streaming Random Forest classifiers.....

2 Related Work

Acquisition and Reconstruction. Capturing the geometry of the surrounding world has been a long standing problem [1]. An offline processing of multiple images allowed to reconstruct digital heritage and construct world-scale 3D models with remarkable quality [6]. The inception of low-cost RGBD sensors and powerful GPUs enabled online 3D scanning, augmented reality or using 3D environment models for navigation purposes [7]. [3] enables 3D reconstruction of small scenes using a single off-the-shelf RGB camera. It uses a sparse tracking method to first estimate the camera's pose and then select key frames and relative to them secondary frames, from which 3D stereo reconstruction is performed. The achieved results are similar to KinectFusion with the only limitation being that the precision of texture-less surfaces' reconstruction is somewhat lacking.

Scene Understanding. Object recognition, detection and segmentation has been done with 2D RGB images, RGBD data, point clouds and volumetric representations. [2] uses a Voxel-based CRF for simultaneous reconstruction and segmentation. Each voxel contains information about visibility and occlusion as well as group membership. The first two are used to to improve reconstruction by mitigating depth-map noise. The visibility values are constrained by that each ray from the camera can hit only one visible vortex. The group membership information encodes priors given by bounding boxes of detected objects. A graph-cut algorithm is used for global inference. The whole scene has to be registered beforehand, since any change in the CRF's structure would require restarting the inference algorithm. No computational performance was reported. [4] is a first step towards online simultaneous registration and segmentation. Using RGBD images, the framework constructs a model of the environment and updates it each time a new frame comes in. When a significant change in the model is detected, the resulting model is split into a static and a dynamic part, where the latter is assumed to have moved in space. The movement is detected by comparing the expected and observed intensity values at each voxel. The system is online, but is far from real time with 0.7 to 2s processing time per frame. [5] shows that converting to another data representation can help to improve segmentation accuracy and inference speed. TSDF model is used to reconstruct a 3D scene from a sequence of RGBD images. The volumetric representation is triangulated and a 3D mesh is recovered. Next, visual features are computed directly on images and projected into the 3D model, while geometric features are computed on the mesh directly. Segmentation is done via a CRF. The approach is tested on the augmented KITTI and NYU datasets for indoor and outdoor scene segmentation where it delivers state-of-the-art results.

References

1. ROBERTS , L. G. 1963. Machine perception of three-dimensional solids. Ph.D. thesis, Massachusetts Institute of Technology.
2. KIM , B.-S., KOHLI , P., AND SAVARESE , S. 2013. 3D scene understanding by voxel-CRF. In Proc. ICCV.
3. PRADEEP , V., RHEMANN , C., IZADI , S., ZACH , C., BLEYER , M., AND BATHICHE , S. 2013. Monofusion: Real-time 3D reconstruction of small scenes with a single web camera. In Proc. ISMAR.
4. HERBST , E., HENRY , P., AND FOX , D. 2014. Toward online 3-d object segmentation and mapping. In IEEE International Conference on Robotics and Automation (ICRA).
5. VALENTIN , J. P., SENGUPTA , S., WARRELL , J., SHAHROKNI , A., AND TORR , P. H. 2013. Mesh based semantic modelling for indoor and outdoor scenes. In Proc. CVPR.
6. LEVOY , M., PULLI , K., CURLESS , B., RUSINKIEWICZ , S., KOLLER , D., EREIRA , L., GINZTON , M., ANDERSON , S., DAVIS , J., G NSBERG , J., ET AL . 2000. The digital Michelangelo project: 3D scanning of large statues. In Proc. SIGGRAPH. ACM.
7. NEWCOMBE , R. A., IZADI , S., HILLIGES , O., MOLYNEAUX , D., KIM , D., DAVISON , A. J., KOHLI , P., SHOTTON , J., HODGES , S., AND FITZGIBBON , A. 2011. KinectFusion: Real-time dense surface mapping and tracking. In Proc. ISMAR.