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 enviornment has never been easier. SemanticPaint can register your surroundings which, after undergoing a low-level 3D reconstruction, can be semanticly 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 instantenous feedback and can re-label any object to correct errors. SemanticPaint makes capturing customized enviornment models with object classes particular to the user's interest easy and efficient.

The pipeline starts with capturing the enviornment 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 volumentric data in the TSDF format directly in order to avoding the costly conversions to mesh or point-cloud formats. To allow this, the Voxel Oriented Patch features — a new type of a descriminative 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 enviornments in robot navigation or to aid people with impaird sight and (3) map enviornments for use in augmanted reality scenerios or games.

The rest of the paper is organised 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 outlint the Streaming Random Forest classifiers.....

2 Related Work

Capturing the geometry of the surrounding world has been a long standing problem. We have managed to reconstruct digital heritage and construct world-scale 3D models with remarkable quality using offline processing from multiple images. Since low-cost RGBD sensors and powerful GPUs have become available, whch enabled online 3D scanning, augmented reality or using 3D enviornment models for navigation purposes.

There are a number of approaches to scene understanding. Some of them work on 2D RGB images, other reconstruct geometry from multiple RGB images but still use only 2D data for classification purposes and back-project the results into the 3D model. Yet other approaches focus on geometric data exclusively. Since RGBD sensors are available there has been a growing interest in working with RGBD data, point clouds or voxel representations. There has been some work on segmenting 3D scenes and meshes, detect objects in small scenes or replace them with synthetic models.

There has been some work on semantic segmentation of 3D meshes; These methods consider mostly noise-free meshes, do not work in real time and use geometric features only. Some work has been done on matching scan data with databases of synthetic 3D models. It would allow to replace noisy point clouds with detailed models to improve reconstruction quality and decrease memory requirements by exploiting repetetiveness. Usually a model is split into parts individually matched against an offline-built database. It is too slow for real-time usage.

There exists an online SLAM system that recognizes objects and perform online model updatres, but it supports only a single object class and relatively small scenes. What's appealing is that it does semantic recognition and reconstruction.

There are no approaches that can handle an outdoor setting. Most of the work focused on image classification. No system works in real-time nor does online model updates. They also do not use ful 3d information, or do so in a global optimization setting which is slow.

[6] enables 3D reconstruction of small since 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 perfomed. Results achieved are similar to KinectFusion with the only limitation being low precision of reconstruction of textureless surfaces.

[7] is a first step towards online simultaneous registration and segmentation. Using RGBD images, the framework constructs a model of the enviornment 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.

[8]

[5] uses a Voxel-based CRF for simultaneous segmentation and reconstruction. 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 inference, which is performed globally. No computational performance was reported.

3 Pipeline

4 Mean-Field Inference

5 Streaming Random Forest

6 Voxel-Oriented Patch Features

7 Qualitative Results

8 Quantitative Results

9 Discussion

10 Concolusions

References

1. J. Hagenauer, E. Offer, and L. Papke. Iterative decoding of binary block and convolutional codes. *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 429-445, Mar. 1996.
2. T. Mayer, H. Jenkac, and J. Hagenauer. Turbo base-station cooperation for intercell interference cancellation. *IEEE Int. Conf. Commun. (ICC)*, Istanbul, Turkey, pp. 356–361, June 2006.
3. J. G. Proakis. *Digital Communications*. McGraw-Hill Book Co., New York, USA, 3rd edition, 1995.
4. F. R. Kschischang. Giving a talk: Guidelines for the Preparation and Presentation of Technical Seminars. http://www.comm.toronto.edu/frank/guide/guide.pdf.
5. KIM , B.-S., KOHLI , P., AND SAVARESE , S. 2013. 3D scene understanding by voxel-CRF. In Proc. ICCV.

6. 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.
7. 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).
8. 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.