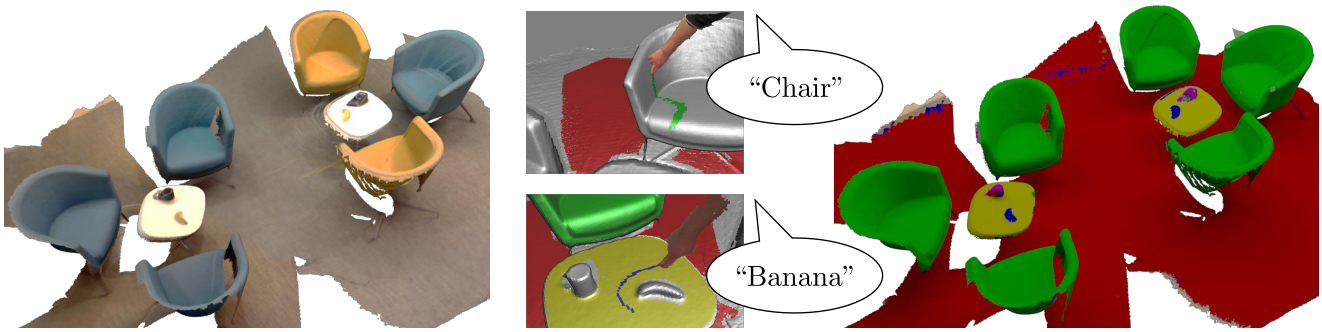
SemanticPaint: Interactive Segmentation and Learning of 3D Worlds

Stuart Golodetz* Michael Sapienza* Julien P. C. Valentin Vibhav Vineet Ming-Ming Cheng
Victor A. Prisacariu Olaf Kähler Carl Yuheng Ren Anurag Arnab Stephen L. Hicks *†
David W. Murray Shahram Izadi Philip H. S. Torr



Introduction We present a real-time, interactive system for the geometric reconstruction, object-class segmentation and learning of 3D scenes [Valentin et al. 2015]. Using our system, a user can walk into a room wearing a depth camera and a virtual reality headset, and both densely reconstruct the 3D scene [Newcombe et al. 2011; Nießner et al. 2013; Prisacariu et al. 2014]) and interactively segment the environment into object classes such as 'chair', 'floor' and 'table'. The user interacts *physically* with the real-world scene, touching objects and using voice commands to assign them appropriate labels. These user-generated labels are leveraged by an online random forest-based machine learning algorithm, which is used to predict labels for previously unseen parts of the scene. The predicted labels, together with those provided directly by the user, are incorporated into a dense 3D conditional random field model, over which we perform mean-field inference to filter out label inconsistencies. The entire pipeline runs in real time, and the user stays 'in the loop' throughout the process, receiving immediate feedback about the progress of the labelling and interacting with the scene as necessary to refine the predicted segmentation.

Background In this demo, we present an interactive approach to the exciting problem of real-time 3D scene segmentation, building on a large body of recent work in geometric scene reconstruction and scene understanding to showcase a system that can allow a user to segment an entire room in a very short period of time. Since we keep the user in the loop, our system can be used to produce high-quality segmentations of a real-world environment. Such segmentations have numerous uses, e.g. (i) we can use them to identify walkable surfaces in an environment as part of the process of generating a navigation map that can provide routing support to people or robots; (ii) we can use them to help partially-sighted people avoid collisions by highlighting the obstacles in an environment; (iii) in a computer vision setting, we can extract 3D models from them that can be used to train object detectors.

*S. Golodetz and M. Sapienza assert joint first authorship. Email: {stuart.golodetz,michael.sapienza}@eng.ox.ac.uk.

†SG, MS, JPCV, VAP, OK, CYR, AA, DWM, SLH and PHST are with the University of Oxford, UK. VV is with Stanford University. MMC is with Nankai University. SI is with Microsoft Research.

SIGGRAPH 2015 Emerging Technologies, August 09 – 13, 2015, Los Angeles, CA.
ACM 978-1-4503-3635-2/15/08.
http://dx.doi.org/10.1145/2782782.2792488

System Pipeline Our system is built on top of a dense 3D reconstruction pipeline, allowing real-time fusion of noisy depth maps into an implicit volumetric surface representation. Our system then allows the user to walk up to any object of interest, simply touch and *'paint'* the physical surface, and vocally call out a new or existing object class name. Our method first cleanly segments any touched object from its supporting or surrounding surfaces, using a new volumetric inference technique based on an efficient mean-field approximation. In the background, a new form of *streaming random forest* is trained and updated as new object examples are labelled by the user. The random forest can quickly infer the likelihood that any newly-observed voxel belongs to each object class. The final stage of our pipeline estimates a spatially-consistent dense labelling of the voxel reconstruction by again performing mean-field inference over the voxel space, but now using the results from the decision forest. This creates a system that can be used to rapidly segment 3D scenes and learn from these labels in an online manner. Furthermore, the user can quickly relabel parts of the scene as necessary and see the improved results almost instantaneously as the learned models are updated online.

Demo We have constructed a small desk environment to support our demo. Visitors will be able to interactively segment this scene in real time and visualise the results on a virtual reality headset. They will specify labels using voice commands and touch interaction. As shown in the supplementary video and demonstrated live at the event, a reasonably-sized room can be reconstructed and labelled in a matter of minutes, with high-quality results, and in a highly-interactive manner. We will be making our framework open-source after the conference so that others can build on our work.

References

NEWCOMBE, R. A. *et al.* 2011. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *ISMAR*, IEEE.

NIESSNER, M. *et al.* 2013. Real-time 3D Reconstruction at Scale using Voxel Hashing. *ACM TOG 32*, 6, 169.

PRISACARIU, V. A., KÄHLER, O. *et al.* 2014. A Framework for the Volumetric Integration of Depth Images. *ArXiv e-prints*.

VALENTIN, J. P. C. *et al.* 2015. SemanticPaint: Interactive 3D Labeling and Learning at your Fingertips. To appear in *ACM TOG*.