Discrete Optimization of Ray Potentials for Semantic 3D Reconstruction

Nikolay Savinov[1], Ľubor Ladický[1], Christian Häne, Marc Pollefeys
Institute for Visual Computing, ETH Zürich

A particularly powerful approach to the problem of dense 3D reconstruction from images is to pose it as a volumetric labeling problem. The volume is segmented into occupied and free space (the inside and the outside of an object) and the surface is extracted as the boundary in between. The data costs are extracted from the input images either directly by computing matching scores per voxel or by first computing depth maps and deriving a per pixel unary potential based on the depth maps. Unary terms, approximately modelling the likelihood the depth for a given pixel agrees with the estimate, encourage voxels in an interval just before the matched 3D point to take the free-space label and voxels in an interval right after the matched 3D point to take the foreground label. However, this assumption does not hold in general. Failures due to this problem lead to blowing up corners, roofs of buildings or thin objects. Another problem is that the unary potential does not model, whether the voxel is visible. The standard approach is also not suitable for incorporating multiple candidate matches along the viewing ray in the optimization together. Furthermore, most methods have been designed for 3D reconstruction without semantic classes.

We propose to formulate an optimization problem which measures the data fidelity directly in image space while still having all the benefits of a volumetric representation. The main idea is to use a volumetric representation, but describe the data cost as a potential over rays. Traversing along a ray from the camera center we observe free space until we first hit an occupied voxel of a certain semantic class and we cannot assume anything about the unobserved space behind. The potential we introduce correctly assigns for each ray the cost, based on the depth and semantic class of the first occupied voxel along the ray. We are interested in finding the smooth solution, whose projection into each camera agrees with the depth and semantic observations. Thus, the energy will take the form:

$$E(\mathbf{x}) = \sum_{r\in\mathcal{R}} \psi_r(\mathbf{x}^r) + \sum_{(i,j)\in\mathcal{E}} \psi_p(x_i, x_j), \tag{1}$$

where each $x_i \in \mathcal{L}$ is the voxel variable taking a label from the label set $\mathcal{L}$ with a special label $l_f \in \mathcal{L}$ corresponding to free space; $\mathcal{R}$ is the set of rays, $\psi_r(.)$ is the ray potential over the set of voxels $\mathbf{x}^r$, $\mathcal{E}$ is the set of local voxel neighbourhoods, and $\psi_p(.)$ is a smoothness enforcing pairwise regularizer. Each ray r of length N_r consists of voxels $x_i^r = x_{r_i}$, where $i \in \{0,1,..N_r-1\}$. The ray potential takes the cost depending only on the first non-free space voxel along the ray. For a 2-label problem, the potential takes the form:

$$\psi_r(\mathbf{x}^r) := \begin{cases} \phi_r(\min(i|x_i^r \neq l_f)) & \text{if } \exists x_i^r \neq l_f \\ \phi_r(N_r) & \text{otherwise,} \end{cases} \tag{2}$$

where N_r is the length of the ray, $\phi_r(i)$ is the cost taken, if i is the first foreground pixel along the ray, and $\phi_r(N_r)$ is the cost for the whole ray being free space. We propose a solution using QPBO relaxation [1], where the energy $E(\mathbf{x})$ is transformed to a submodular energy $E(\mathbf{x},\overline{\mathbf{x}})$ with additional variables $\overline{x_i} = 1 - x_i$, and solved by relaxing these constraints. To make our problem solvable using graph-cut, our goal is to transform these potentials into a pairwise energy with additional auxiliary variables $\mathbf{z}$, such that:

$$\psi_r(\mathbf{x}^r) = \min_{\mathbf{z}} \psi_q(\mathbf{x}^r, \overline{\mathbf{x}}^r, \mathbf{z}), \tag{3}$$

where $\psi_q(.)$ is pairwise submodular. Additionally, to keep the problem feasible, we find a transformation, for which the number of edges in the graph with auxiliary variables grows at most linearly with length of a ray. To achieve this goal we perform these five steps:

1. Polynomial representation of the ray potential,

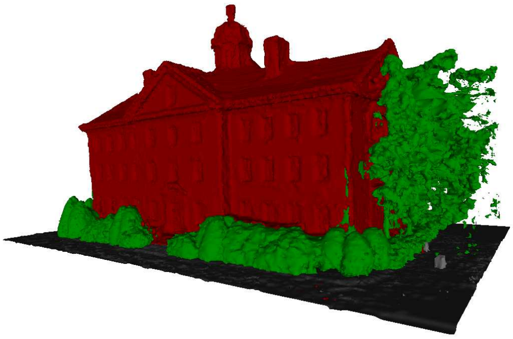

Figure 1: *Our method successfully reconstructed challenging 3D data with high level of detail. Minor errors in the reconstructions were caused by the combination of errors of the semantic classifier, insufficient amount data from certain viewpoints or errors in the depth prediction for smooth textureless surfaces.*

2. Transformation into higher order submodular potential using additional variables $\overline{\mathbf{x}}$,
3. Pairwise graph construction of a higher order submodular potential using auxiliary variables $\mathbf{z}$,
4. Merging variables [5] to get the linear dependency of the number of edges on length,
5. Transformation into a normal form, symmetric over $\mathbf{x}$ and $\overline{\mathbf{x}}$, suitable for QPBO [1].

Following the α-expansion [2] approach, we decompose the multi-label optimization problem into a set of 2-label problems, where in each subproblem every variable can either keep its old label or change it to an expanded label α. Each subproblem takes the same form as the general 2-label formulation, and thus can be solved approximately using the proposed QPBO relaxation.

As an input our method uses semantic likelihoods, predicted by a pixelwise context-based classifier from [4], and top 3 depth likelihoods obtained by a plane sweep stereo matching algorithm using zero-mean normalized cross-correlation. As a smoothness enforcing pairwise potential we used the discretized anisotropic pairwise regularizer [3].

We tested our algorithm on 6 datasets - South Building, Catania, CAB, Castle-P30, Providence and Vienna Opera. Our method managed to successfully reconstruct all 3D scenes with a relatively high precision. Our method managed to fix systematic reconstruction artifacts, caused by approximations in the modelling of the true ray likelihoods.

[1] E. Boros and P.L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 2002.

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[3] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[4] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *International Conference on Computer Vision*, 2009.

[5] Srikumar Ramalingam, Chris Russell, Lubor Ladicky, and Philip HS Torr. Efficient minimization of higher order submodular functions using monotonic boolean functions. *Arxiv preprint arXiv:1109.2304*, 2011.

[1] The authors assert equal contribution and joint first authorship