Joint 3D Scene Reconstruction and Class Segmentation

Christian Häne[1], Christopher Zach[2], Andrea Cohen[1], Roland Angst[1*], Marc Pollefeys[1]

[1]ETH Zürich, Switzerland
{chaene, acohen, rangst, pomarc}@inf.ethz.ch

[2]Microsoft Research Cambridge, UK
chzach@microsoft.com

Abstract

Both image segmentation and dense 3D modeling from images represent an intrinsically ill-posed problem. Strong regularizers are therefore required to constrain the solutions from being 'too noisy'. Unfortunately, these priors generally yield overly smooth reconstructions and/or segmentations in certain regions whereas they fail in other areas to constrain the solution sufficiently. In this paper we argue that image segmentation and dense 3D reconstruction contribute valuable information to each other's task. As a consequence, we propose a rigorous mathematical framework to formulate and solve a joint segmentation and dense reconstruction problem. Image segmentations provide geometric cues about which surface orientations are more likely to appear at a certain location in space whereas a dense 3D reconstruction yields a suitable regularization for the segmentation problem by lifting the labeling from 2D images to 3D space. We show how appearance-based cues and 3D surface orientation priors can be learned from training data and subsequently used for class-specific regularization. Experimental results on several real data sets highlight the advantages of our joint formulation.

1. Introduction

Even though remarkable progress has been made in recent years, both image segmentation and dense 3D modeling from images remain intrinsically ill-posed problems. The standard approach to address this ill-posedness is to regularize the solutions by introducing a respective prior. Traditionally, the priors enforced in image segmentation approaches are stated entirely in the 2D image domain (e.g. a contrast-sensitive spatial smoothness assumption), whereas priors employed for image-based reconstruction typically yield piece-wise smooth surfaces in 3D as their solutions. In this paper we demonstrate that joint image segmentation and dense 3D reconstruction is beneficial for both tasks. While the advantages of a joint formulation for segmentation and 3D reconstruction have already been observed and utilized in the literature, our main contribution is the introduction of a rigorous mathematical framework to formulate and solve this joint optimization task. We extend volumetric scene reconstruction methods, which segment a volume of interest into occupied and free-space regions, to a multi-label volumetric segmentation framework assigning object classes or a free-space label to voxels. On the one hand, such a joint approach is highly beneficial since the associated appearance (and therefore a likely semantic category) of surface elements can influence the spatial smoothness prior. Thus, a class-specific regularizer guided by image appearances can adaptively enforce spatial smoothness and preferred orientations of 3D surfaces. On the other hand, densely reconstructed models induce image segmentations which are guaranteed to correspond only to geometrically meaningful objects in 3D. Hence, the segmentation results are trivially consistent across multiple images.

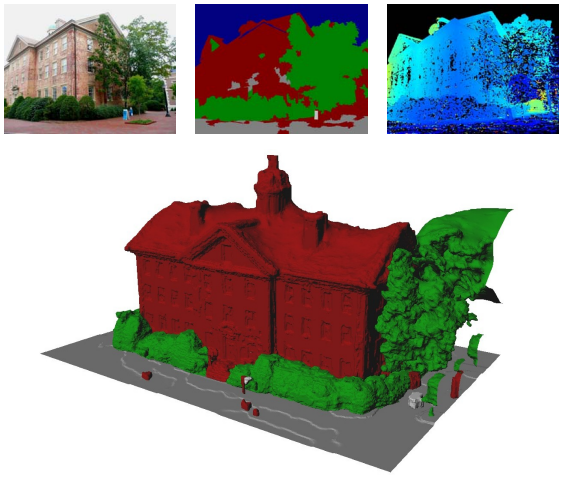

Figure 1: Top: Example of input image, standard image classification result, depthmap. Bottom: Our proposed joint optimization combines class segmentation and geometry resulting in an accurately labeled 3D reconstruction

In a nutshell, we propose to learn appearance likelihoods and class-specific geometry priors for surface orientations

*Now at Stanford University

from training data in an initial step. These data-driven priors can then be used to define unary and pairwise potentials in a volumetric segmentation framework, complementary to the measured evidence acquired from depth maps. While optimizing over the label assignment in this volume, the image-based appearance likelihoods, depth maps from computational stereo, and geometric priors interact with each other yielding an improved dense reconstruction and labeling. The remainder of the paper explains each step in detail, and our mathematical framework is verified on several challenging real-world data sets.

2. Related Work

There is a vast literature on dense 3D modeling from images. Here we sketch only a small subset of related literature, and refer e.g. to the Middlebury MVS evaluation page [20] for a broader survey. Given a collection of depth images (or equivalently densely sampled oriented 3D points) the methods proposed in [13, 27, 23] essentially utilize the surface area as regularization prior, and obtain the final surface representation indirectly via volumetric optimization. One main difference between [13] and [27, 23] is the utilization of a combinatorial graph-cut formulation in the former, whereas [27, 23] employ a continuously inspired numerical scheme. The regularization prior in these works is isotropic, i.e. independent of the surface normal (up to the impact of the underlying discretization), corresponding to a total variation (TV) regularizer in the volumetric representation. The work of [10] utilizes an anisotropic TV prior for 3D modeling in order to enforce the consistency of the surface normals with a given normal field, thus better preserving high frequency details in the final reconstruction. All of the above mentioned work on volumetric 3D modeling from images returns solely a binary decision on the occupancy state of a voxel. Hence, these methods are unaware of typical class-specific geometry, such as the normals of the ground plane pointing upwards. These methods are therefore unable to adjust the utilized smoothness prior in an object- or class-specific way. This observation led to the initial motivation for the present work. More specifically, it is notoriously difficult to faithfully reconstruct weakly or indirectly observed parts of the scene such as the ground, which is usually captured in images at very slanted angles (at least in terrestrial image data). [9] proposes to extend an adaptive volumetric method for surface reconstruction in order not to miss important parts of the scene in the final geometry. The assumption in their method is that surfaces with weak evidence are likely to be real surfaces if adjacent to strongly observed freespace. A key property of our work is that weakly supported scene geometry can be assisted by a class-specific smoothness prior.

If only a single image is considered and direct depth cues from multiple images are not available, assigning object categories to pixels yields crucial information about the 3D scene layout [8, 19], e.g. by exploiting the fact that building facades are usually vertical, and ground is typically horizontal. These relations are generally not manually encoded, but extracted from training data. Such known geometric relations between object categories can also be helpful for 2D image segmentation, e.g. by assuming a particular layout for indoor images [15], a tiered layout [6] or class-specific 2D smoothness priors [22]. Utilizing appearance-based pixel categories and stereo cues in a joint framework was proposed in [11] in order to improve the quality of obtained depth maps and semantic image segmentations. In our work, we also aim on joint estimation of 3D scene geometry and assignment of semantic categories, but use a completely different problem representation—which is intrinsically using multiple images—and solution method. [18, 2] also present joint segmentation and 3D reconstruction methods, but the determined segments correspond to individual objects (in terms of an underlying *smooth* geometry) rather than to semantic categories. Furthermore, a method [1] using semantic information for dense object reconstruction in form of shape priors has been developed concurrently to our work.

3. Joint 3D Reconstruction and Classification

In this section we describe the underlying energy formulation for our proposed joint surface reconstruction and classification framework and its motivation. Similar to previous works on global surface reconstruction we lift the problem from an explicit surface representation to an implicit volumetric one. The increased memory consumption is compensated by the advantages of allowing arbitrary but closed and oriented topology for the resulting surface.

3.1. Continuous Formulation

We cast the ultimate goal of semantically guided shape reconstruction as a volumetric labeling problem, where one out of $L+1$ labels is assigned to each location $z \in \Omega$ in a continuous volumetric domain $\Omega \subset \mathbb{R}^3$. In the following we will use indices i and j for labels. Allowed labels are "free/empty space" (with numeric value 0) and "occupied space" with an associated semantic category (values from $\{1, \dots, L\}$). The label assignments will be encoded with $L+1$ indicator functions $x^i : \Omega \to [0,1]$, $i \in \{0, \dots, L\}$: $x^i(z) = 1$ iff label i is assigned at $z \in \Omega$. Note that in the following, the dependence of all quantities on the 3D location z will be indicated with a subscript to be more consistent with the later discrete formulation, i.e. $x^i(z) = x^i_z$. With this notation in place, the convex relaxation of the labeling problem in a continuous volumetric domain Ω reads as

$$E_{\text{cont}}(\mathbf{x}, \mathbf{y}) = \int_\Omega \sum_i \rho^i_z x^i_z + \sum_{i,j:i<j} \phi^{ij}_z(y^{ij}_z)\, dz, \qquad (1)$$

where $y^{ij} : \Omega \to [-1,1]^3$, $i \in \{0,\dots,L\}$ with $j > i$ are "jump processes" satisfying a modified marginalization constraint

$$\nabla_z x^i = \sum_{j:j>i} y^{ij} - \sum_{j:j<i} y^{ji}. \tag{2}$$

$\rho^i : \Omega \to \mathbb{R}$ encodes the local preference for a particular label. Note that the smoothness term in Eq. 1 is an extension of the standard length/area-based boundary regularizer to Finsler metrics (see e.g. [16]) and the infinitesimal *length functions* $\phi_z^{ij} : \mathbb{R}^3 \to \mathbb{R}_0^+$ are naturally extended from $\mathbb{S}^2$ to $\mathbb{R}^3$, rendering ϕ_z^{ij} a convex and positively 1-homogeneous function. Such choice of ϕ_z^{ij} generalizes the notion of total variation to location and orientation dependent penalization of segmentation boundaries. In addition to the marginalization constraints in Eq. 2, the functions x^i also need to satisfy the normalization constraint, $\sum_i x^i \equiv 1$, and non-negativity constraints. We refer to [25] for a detailed derivation and theoretical interpretation of this energy. A minimizer $(\mathbf{x}, \mathbf{y})$ induces a partition of Ω into free space and respective object categories. The boundaries between the individual regions form the 3D surfaces of interest.

3.2. Discretized Formulation

A disadvantage of this *continuous energy formulation* is that the class of smoothness priors ϕ_s^{ij} is restricted to metrics under reasonable assumptions (see e.g. [12]). Consequently, we focus our attention on discrete lattices (i.e. regular voxel grids) as underlying domain where these restrictions do not apply. Hence, Ω denotes a finite voxel grid with voxels $s \in \Omega$ in the following. A discrete version of the continuous energy in Eq. 1 not requiring a metric prior reads as [24]

$$E_{\text{discr}}(\mathbf{x}) = \sum_{s\in\Omega}\Big(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(x_s^{ij} - x_s^{ji})\Big) \tag{3}$$

subject to the following marginalization, normalization and non-negativity constraints,

$$\begin{aligned} &x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k \quad (k \in \{1,2,3\}) \\ &x_s \in \Delta, \qquad x_s^{ij} \geq 0. \end{aligned} \tag{4}$$

$e_k \in \mathbb{R}^3$ denotes the k-th canonical basis vector and $(\cdot)_k$ is the k-th component of its argument. The discrete marginalization constraints above follow from Eq. 2 by employing a forward finite difference scheme for the spatial gradient. The probability simplex of appropriate dimension is denoted by Δ. The variables appearing in Eq. 3 have the following interpretation in the context of joint surface reconstruction and segmentation tasks:

- x_s^i encodes whether label i (i.e. free space or one of the solid object categories) is assigned at voxel s,
- $x_s^{ij} - x_s^{ji} \in [-1,1]^3$ represents the local surface orientation if it is non-zero,
- ρ_s^i is the unary data term encoding the measured evidence, i.e. the preference of voxel s for a particular label i. This data term captures the evidence from two sources: firstly, the measurements from a set of depth maps, and secondly, appearance-based classification scores from the input images as obtained from previously trained classifiers. Section 4 describes in detail how this unary term is modeled.
- Finally, ϕ_s^{ij} is the location and direction-dependent smoothness prior indicating the local compatibility of a boundary between label i and j. Hence, these priors encode the previously mentioned class-specific geometric priors. Of highest importance is the directly observable boundary between free space and any of the object categories. Modeling ϕ_s^{ij} from training data is explained in Section 5.

We will restrict ourselves to homogeneous priors in the following, i.e. the local smoothness contribution $\phi_s^{ij}(x_s^{ij} - x_s^{ji})$ does not depend on s, and the objective in Eq. 3 slightly simplifies to

$$E_{\text{discr}}(\mathbf{x}) = \sum_{s\in\Omega}\Big(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi^{ij}(x_s^{ij} - x_s^{ji})\Big). \tag{5}$$

The rationale behind the spatial homogeneity assumption is that only the orientation of a boundary surface and the affected labels are of importance, but not the precise location.

Once the values of ρ_s^i are determined and the smoothness priors ϕ^{ij} are known, the task of inference is to return an optimal volumetric labeling. Since we employ a convex problem stated in Eq. 3, any convex optimization suitable for non-smooth programs can be utilized. After introducing Lagrange multipliers for the constraints and after biconjugation of the smoothness term we are able to directly apply the primal-dual algorithm proposed in [4]. We briefly outline the numerical scheme used in our experiments in the supplementary material.

4. The Ray Likelihood and Its Approximation

In this section we describe how available depth maps (with potentially missing depth values) and appearance-based class likelihoods are converted into respective unaries ρ for joint volumetric reconstruction and classification as described in the previous section. A completely sound graphical model relating image observations with occupancy states of 3D voxels requires observation likelihoods corresponding to clique potentials with entire rays in 3D forming cliques (e.g. [14]). In the following we argue that—under suitable smoothness assumptions on the solution—we can approximate the higher-order clique potentials by suitable unary

ones. We aim on factorizing the clique potential into only unary terms such that the induced (higher-order) cost of a particular boundary surface is approximated by the unaries. Additionally we employ the usual assumption of independence of observations across images. This means, that the unary potentials described below based on color images (and associated depth maps) are accumulated over all images to obtain the effective unary potentials.

In the following we consider a particular pixel p in one of the input images (respectively depth maps, since we assume that depth images use color images as their reference views). The pixel p induces a ray in 3D space, which leads to a set of traversed voxels $s \in \text{ray}(p)$ and the corresponding latent variables x_s^i and their associated unary potentials ρ_s^i. Recall that i indexes one of the $L+1$ semantic categories $\{0, 1, \ldots, L\}$ with 0 corresponding to sky (i.e. free space) and i indicating object category i, respectively. Our task is to (approximately) model the likelihoods

$$P\left(\hat{d}(p), \hat{A}(p) \mid \{x_s^i\}_{s \in \text{ray}(p)}\right), \tag{6}$$

where $\hat{d}(p)$ is the observed depth at pixel p (which may be missing), and $\hat{A}(p)$ encapsulates the local image appearance in the neighborhood of p. Note that in terms of a graphical model the respective potential, $-\log P\big(\hat{d}, \hat{A} \mid \{x_s^i\}_{s \in \text{ray}(p)}\big)$, depends on the entire clique $\{s : s \in \text{ray}(p)\}$. Clearly, for a particular ray the likelihoods of observing $\hat{d}(p)$ and $\hat{A}(p)$ only depend on the first crossing from freespace to occupied space. Nevertheless, proper handling of voxel visibility links all voxels along the ray to form a clique.

For notational convenience we will drop the dependence on the pixel p, and also index voxels along ray(p) by their depth with respect to the current view. We will substantially simplify the potentials (and therefore the inference task) by considering the following cases:

Observed depth: This is the case when $\hat{d}$ in the depth map is valid (i.e. not missing). In this case we assume that Eq. 6 factorizes into

$$P\big(\hat{d}, \hat{A} \mid \text{voxel } d \text{ is first crossing to } i\big) = P(\hat{d} \mid d) P(\hat{A} \mid i),$$

where $P(\hat{d} \mid d)$ captures the noise for inliers in the depth sensing process and is usually a monotonically decreasing function of $|d - \hat{d}|$. $P(\hat{A} \mid i)$ is induced by the confidence of an appearance-based classifier for object category i.

We only define non-zero unaries for voxels along the ray near the observed depth. Assume that the inlier noise of depth estimation is bounded by δ, and we denote by $\hat{d} \pm \delta$ the voxels along the ray with distance δ and $-\delta$, respectively. We set the unary potentials

$$\rho^i_{\hat{d}+\delta} = \sigma_{\text{class } i} \qquad \rho^i_d = \begin{cases} 0 & \text{for } i = 0 \\ \eta(\hat{d} - d) & \text{for } i > 0. \end{cases} \tag{7}$$

for voxels d near the observed depth, i.e. voxels d closer to $\hat{d}$ than δ. Here $\sigma_{\text{class } i} = -\log P(\hat{A} \mid i)$ (with class 0 corresponding to sky). The function $\eta : [-\delta, \delta] \to \mathbb{R}$ is independent of the object category i and reflects the noise assumptions of $\hat{d}$. We choose $\eta(\hat{d} - d) = \beta \operatorname{sgn}(\hat{d} - d)$ for $\beta > 0$, corresponding to an exponentially distributed noise for depth inliers. Inserting unaries only near the observed depth corresponds to truncating the cost function, hence we assume exponentially distributed inliers and uniformly distributed outlier depth values. See Fig. 2 for an illustration of unaries along the ray.

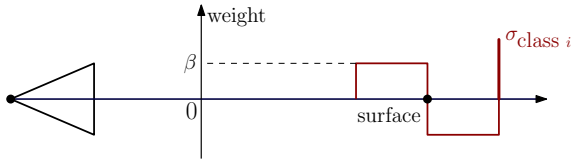

Figure 2: Unaries assigned to voxels along a particular line-of-sight.

Since we enforce spatial smoothness of the labeling (i.e. multiple crossings within the narrow band near $\hat{d}$ are very unlikely), we expect three possible configurations for voxels in $[\hat{d} - \delta, \hat{d} + \delta]$ described below. For each configuration we state the contribution of unary terms for the particular ray to the complete energy.

1. In the labeling of interest we have that free-space transitions to a particular object class i at depth d. Hence, $x_s^0 = 1$ for $s \in [\hat{d} - \delta, d)$ and $x_s^i = 1$ for $s \in [d, \hat{d} + \delta]$. Summing the unaries according to Eq.7 over the voxels in $[\hat{d} - \delta, \hat{d} + \delta]$ yields

$$\sigma_{\text{class } i} + \sum_{d' \in [d, \hat{d}+\delta]} \eta(\hat{d} - d'),$$

   i.e. the negative log-likelihood of observing the appearance category i in the image and the one corresponding to the depth noise assumption. Note that the second term, $\sum_{d'} \eta(\hat{d} - d')$ will be non-positive and therefore lower the overall energy. This beneficial term is not appearing in the other cases below.

2. If all voxels in the particular range $[\hat{d} - \delta, \hat{d} + \delta]$ are freespace ($x_s^0 = 1$ for all the voxels in this range), then the contribution to the total energy is just σ_{sky}. Since a potential transition to a solid object class outside the near band is not taken into account, this choice of unary potentials implicitly encodes the assumption that that freespace near the observed depth implies freespace along the whole ray.

3. All voxels in the range are assigned to object label i (i.e. $x_s^i = 1$ for $s \in [\hat{d} - \delta, \hat{d} + \delta]$. This means that there

was a transition from freespace to object type i earlier along the ray. Thus, the contribution to the energy is $\sigma_{\text{class } i}$ in this case.

Overall, our choice of unaries will faithfully approximate the desired true data costs in most cases. Since camera centers are in free-space by definition, we add a slight bias towards free-space along the line-of-sight from the respective camera center to the observed depth (i.e. voxels in the range $[0, \hat{d} - \delta]$). This has also a positive influence on the convergence speed.

Missing depth: If no depth was observed at a particular pixel p, we cannot assign unaries along the corresponding ray. Since missing depth values mostly occur in the sky regions of images, we found the following modification helpful to avoid "bleeding" of buildings etc. beyond their respective silhouettes in the image: in case of missing depth we set the unary potentials to

$$\rho_s^0 = \min\{0, \sigma_{\text{sky}} - \min_{i \neq \text{sky}} \sigma_i\} \tag{8}$$

and $\rho_s^i = 0$ for $i > 0$ for all voxels s along ray(p). This choice of unaries favors freespace along the whole ray whenever depth is missing and sky is the most likely class label in the image.

5. Training the Priors

In this section, we will explain how the appearance likelihoods used in the unary potentials ρ_s^i and the class-specific geometric priors ϕ^{ij} are learned from training data. While the appearance terms are based on classification scores of a standard classifier, training of geometric priors from labeled data is more involved. We first start describing the training of the appearance likelihoods before discussing the training procedure for smoothness priors.

5.1. Appearance Likelihoods

In order to get classification scores for the labels in the input images we train a boosted decision tree classifier [7] on manually labeled training data. In a first step, the training images are segmented into super-pixels using the mean shift segmentation algorithm[1]. Features are extracted for each super-pixel. We use the default parameters as implemented by [7], resulting in 225 dimensional feature vectors based on color, intensity, geometry, texture and location. It should be noted that the geometry and location features are extracted by using 2-D information on the images (superpixel size, shape, and relative position in the image) and they are not related to the 3-D geometry of the scene. The extracted features and ground truth annotations are fed into the boosted decision tree. The classifier is trained over 5 classes: sky, building, ground, vegetation, and clutter. We use 2 splits per decision tree and 200 boosting rounds. We designed a training dataset by taking 76 images from the CamVid dataset [3] and 101 images from the MSRC dataset. We also added 34 images taken at street level of different buildings. These buildings are not part of the evaluation data set.

Once the classifier is trained, it can be used to obtain scores for each region of the input images. These scores represent the log-likelihoods of each class for each region of the image.

[1]OpenCV implementation

5.2. Class-Specific Geometric Priors

We use a parametric model for the functions ϕ_s^{ij} appearing in the smoothness term of Eq. 5. As already mentioned we restrict ourselves to spatially homogeneous functions $\phi_s^{ij} = \phi^{ij}$, and thus there is no dependency on the location s. Note that the energy formulation in Eq. 5 naturally corresponds to a negative log-probability. Hence, the functions ϕ^{ij} will be also interpreted as negative log-probabilities. Let $s_{i \leftrightarrow j}$ denote a transition event between labels i and j at some voxel s, and let n_s^{ij} be the (unit-length) boundary normal at this voxel. Instead of modeling ϕ^{ij} directly, we use

$$P(n_s^{ij}) = P(n_s^{ij} \mid s_{i \leftrightarrow j}) P(i \leftrightarrow j), \tag{9}$$

where we applied the homogeneity assumption, i.e. $P(s_{i \leftrightarrow j}) = P(i \leftrightarrow j)$. The conditional probability, $P(n_s^{ij} \mid s_{i \leftrightarrow j})$ is now modeled as a Gibbs probability measure

$$P(n_s^{ij} \mid s_{i \leftrightarrow j}) = \exp\left(-\psi^{ij}(n_s^{ij})\right) / Z^{ij}, \tag{10}$$

for a pos. 1-homogeneous function ψ^{ij}. Z^{ij} is the respective partition function, $Z^{ij} \stackrel{\text{def}}{=} \int_{n \in \mathbb{S}^2} \exp\left(-\psi^{ij}(n^{ij})\right) \, dn$, and $\mathbb{S}^2$ is the 3-dimensional unit sphere. Consequently, ϕ^{ij} in Eq. 5 is now given by

$$\phi^{ij}(n) = \psi^{ij}(n) + \log Z^{ij} - \log P(i \leftrightarrow j) \tag{11}$$

for a unit vector $n \in \mathbb{S}^2$. Maximum-likelihood estimation is used to fit the parameters to available training data, formally

$$\theta = \arg\max_\theta \prod_s \prod_{i,j} P(n_s^{ij} \mid s_{i \leftrightarrow j}) P(i \leftrightarrow j), \tag{12}$$

where the product goes over all training samples s and ψ^{ij} and Z^{ij} are functions of the parameters θ^{ij} which are gathered in $\theta = \{\theta^{ij} \mid i, j \in \{0, \ldots, L\}\}$. In our implementation, we estimate the discrete probabilities $P(i \leftrightarrow j)$ of observing a transition $i \leftrightarrow j$ upfront by counting the relative frequencies $N^{ij} / \sum_{i,j} N^{ij}$ of the respective type of boundaries from training data. Estimating first $P(i \leftrightarrow j)$ has the advantage that the ML-estimation in Eq. 12 decouples into independent estimation problems of the form

$$\theta^{ij} = \arg\min_{\theta^{ij}} \sum_{k=1}^{N^{ij}} \psi^{ij}(n_k^{ij}; \theta^{ij}) + N^{ij} Z^{ij}(\theta^{ij}), \tag{13}$$

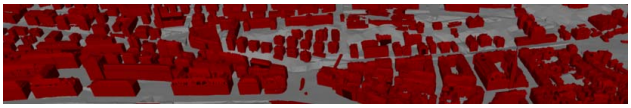

Figure 3: A section of the cadastral city model used to train the geometric priors.

where the summation goes over all the N^{ij} transition samples n_k^{ij} between labels i and j. Since for many choices of ψ^{ij} the partition function cannot be solved analytically, we use Monte Carlo integration to obtain an estimate for Z^{ij}. Given the low dimensionality of θ^{ij} (up to 4 components, see Section 5.3 below) and the necessity of Monte Carlo integration for the partition function, we use a simple grid search to find an approximate minimizer θ^{ij}. As training data we use a three dimensional cadastral city model (see Fig. 3) which enables us to train ψ^{ij} for the transitions ground $\leftrightarrow$ free space, ground $\leftrightarrow$ building and building $\leftrightarrow$ free space. Label transitions unobserved in the training data are defined manually. At this point we need to address two small technical issues:

Remark 1. ϕ^{ij} is only specified for unit vectors $n \in \mathbb{S}^2$, but the argument in the energy model Eq. 5 are usually non-normalized gradient directions $y_s^{ij} \stackrel{\text{def}}{=} x_s^{ij} - x_s^{ji} \in [-1,1]^3$. However, remember that ψ^{ij} is a convex and positively 1-homogeneous function. Together with the fact that the area of the surface element in finite difference discretizations is captured exactly by $\|y_s^{ij}\|_2$, we derive the contribution of y_s^{ij} to the regularizer as

$$\|y_s^{ij}\|_2 \phi^{ij}\left(y_s^{ij}/\|y_s^{ij}\|_2\right) = \phi^{ij}(y_s^{ij})$$

by the 1-homogeneity of ϕ^{ij}. Therefore, the extension of ϕ^{ij} as given in Eq. 11 to arbitrary arguments $y \in \mathbb{R}^3$ is

$$\phi^{ij}(y) = \psi^{ij}(y) + \|y\|_2 \underbrace{\left(\log Z^{ij} - \log P(i \leftrightarrow j)\right)}_{\stackrel{\text{def}}{=} C^{ij}}. \quad (14)$$

Consequently, our smoothness prior ϕ^{ij} will always be composed of an anisotropic, direction-dependent component ψ^{ij} and an isotropic contribution proportional to $C^{ij} = \log Z^{ij} - \log P(i \leftrightarrow j)$. This also implies that there is no need to explicitly model any isotropic component in ψ^{ij}.

Remark 2. The function ϕ^{ij} given in Eq. 14 above is positively 1-homogeneous if ψ^{ij} is, but convexity can only be guaranteed whenever $C^{ij} = \log Z^{ij} - \log P(i \leftrightarrow j) \geq 0$ or $P(i \leftrightarrow j) \leq Z^{ij}$. This is in practice not a severe restriction, since for a sufficiently fine discretization of the domain the occurrence of a boundary surface is a very rare event and therefore $P(i \leftrightarrow j) \ll 1$.

5.3. Choices for ψ^{ij}

We need to restrict ψ^{ij} to be convex and positively 1-homogeneous. One option is to parametrize $\psi^{ij}(n) = \psi^{ij}(n;\theta)$ in the primal and to limit θ such that the resulting ψ^{ij} has these properties, but this may be difficult in general. We choose a slightly different route and parametrize the convex conjugate of ψ^{ij}, $\left(\psi^{ij}\right)^*$,

$$\left(\psi^{ij}\right)^*(p) = \max_n \left\{p^T n - \psi^{ij}(n)\right\} = \imath_{W_{\psi^{ij}}}(p),$$

i.e. the indicator function for a (convex) shape $W_{\psi^{ij}}$ (which will be called a *Wulff shape* [17] in the following). We find it easier to model parametric convex Wulff shapes $W_{\psi^{ij}}$ rather than ψ^{ij} directly. Below we describe the utilized Wulff shapes and its parametrizations. Which Wulff shape is picked for ψ^{ij} (in addition to its continuous parameters) is part of the ML estimation. The description below is for Wulff shapes in a canonical position, since any ψ induced by a rotated shape can be expressed using a canonical one,

$$\psi(n;R) = \max_{p \in R \cdot W_\psi} p^T n = \max_{p \in W_\psi} (Rp)^T n = \psi(R^T n; I).$$

Given remark 1 above there is no need to model the Wulff shape with an isotropic and an anisotropic component (i.e. as Minkowski sum of a sphere and some other convex shape).

The Wulff shapes described below are designed to model two frequent surface priors encountered in urban environments: one prior favors surface normals that are in alignment with a specific direction (e.g. ground surface normals prefer to be aligned with the vertical direction), and the second Wulff shape favors surface normals orthogonal to a given direction (such as facade surfaces having generally normals perpendicular to the vertical direction). In order to obtain a discriminative prior we assume that an approximate vertical direction is provided. We refer to the supplementary material for graphical illustrations of the Wulff shapes and induced smoothness costs.

Line Segment This Wulff shape has only one parameter l and is a line segment in z-direction centered at the origin with length $2l$ (i.e. its endpoints are $(0,0,l)^T$ and $(0,0,-l)^T$). This shape translates to a function $\psi(n) = l|n_3|$, which is convex as long $l \geq 0$.

Half-sphere plus spherical cap This Wulff shape W_ψ consists of a half-sphere with radius r centered at the origin in opposition to a spherical cap with height h. The corresponding function ψ favors directions pointing upwards and isotropically penalizes downward pointing normals. ψ can be computed in closed form (with $n = (n_1, n_2, n_3)^T$),

$$\psi(n) = \begin{cases} r\|n\| & \text{if } n_3 \leq 0 \\ \|n\|\left(\frac{r^2}{2h} + \frac{h}{2}\right) - n_3\left(\frac{r^2}{2h} - \frac{h}{2}\right) & \text{if } (*) \\ r\left\|\binom{n_1}{n_2}\right\| & \text{otherwise,} \end{cases}$$

where $(*)$ is $n_3 > 0$ and $n_3(h^2 + r^2) > (r^2 - h^2)/\|n\|$. By construction W_ψ is convex (and therefore also ψ) as long as $r \geq 0$ and $h \in [0, r]$.

6. Experiments

In this section we present the results obtained on four challenging real world datasets. We compare our geometry to a standard volumetric fusion (in particular "TV-Flux" [23]) and also illustrate the improvement of the class segmentation compared to a single image best-cost segmentation.

We use the dataset castle P-30 from [21] and three additional urban datasets ranging from 127 to 195 images in the dataset size. Camera poses where obtained with the publicly available structure from motion pipeline [26]. The depth maps are computed using plane sweep stereo matching for each of the images with zero mean normalized cross correlation (ZNCC) matching costs. Up to nine images are matched to the reference view simultaneously with best K occlusion handling. To get rid of the noise the raw depth maps are filtered by discarding depth values with a ZNCC matching score above 0.4. The class scores are obtained by using the boosted decision tree classifier explained in Section 5.1. To align the voxel grid with the scene we use the approach described in [5]. We use a multi-threaded C++ implementation to find a minimizer of Eq. 5 (running on a 48 cores).

Fig. 4 illustrates the results for all 4 datasets. As expected, computational stereo in particular struggles with faithfully capturing the ground, which is represented by relatively few depth samples. Consequently, depth integration methods with a generic surface prior such as TV-Flux easily remove the ground and other weakly observed surfaces (due to the well-known shrinking bias of the employed boundary regularizer). In contrast, our proposed joint optimization leads to more accurate geometry, and at the same time image segmentation is clearly improved over a greedy best-cost class assignment.

The third column in Fig. 4 illustrates that the most probable class labels according to the trained appearance likelihoods especially confuses ground, building, and clutter categories. Fusing appearance likelihood over multiple images and incorporating the surface geometry almost perfectly disambiguates the assigned object classes. The joint determination of the right smoothness prior also enables our approach to fully reconstruct ground and all the facades as seen in Fig. 4, 4th column. The ground is consistently missing in the TV-Flux results, and partially the facades and roof structure suffer from the generic smoothness assumption Fig. 4, 5th column). We selected a weighting between data fidelity and smoothness in the TV-Flux method such that successfully reconstructed surfaces have a (visually) similar level of smoothness than the results of our proposed method.

7. Conclusion

We present an approach for dense 3D scene reconstruction from multiple images and simultaneous image segmentation. This challenging problem is formulated as joint volumetric inference task over multiple labels, which enables us to utilize class-specific smoothness assumptions in order to improve the quality of the obtained reconstruction. We use a parametric representation for the respective smoothness priors, which yields a compact representation for the priors and—at the same time—allows to adjust the underlying parameters from training data. We demonstrate the benefits of our approach over standard smoothness assumptions for volumetric scene reconstruction on several challenging data sets.

Future work needs in particular to address the scalability of the method. As a volumetric approach operating in a regular voxel grid, our method shares the limitations in terms of spatial resolution with most other volumetric approaches. Adaptive representations for volumetric data can be a potential solution. We also plan to extend the number of object categories to obtain a finer-grained segmentation. Note that not all pairwise transitions between labels in 3D are equally important or even occur in practice. This fact can be utilized to improve the computational efficiency of our proposed formulation.

Acknowledgements: We thank Carsten Rother and Pushmeet Kohli for initial discussions, and Arnold Irschara for providing dataset 3. Furthermore we gratefully acknowledge the support of the 4DVideo ERC starting grant #210806 and V-Charge grant #269916 both under the EC's FP7/2007-2013. Roland Angst was a recipient of the Google Europe Fellowship.

References

[1] Y. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *Proc. CVPR*, 2013.

[2] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo-joint stereo matching and object segmentation. In *Proc. CVPR*, pages 3081–3088, 2011.

[3] G. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[4] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J. Math. Imag. Vision*, pages 1–26, 2010.

[5] A. Cohen, C. Zach, S. Sinha, and M. Pollefeys. Discovering and exploiting 3d symmetries in structure from motion. In *Proc. CVPR*, 2012.

[6] P. F. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *Proc. CVPR*, pages 3097–3104, 2010.

[7] S. Gould, O. Russakovsky, I. Goodfellow, P. Baumstarck, A. Y. Ng, and D. Koller. The STAIR Vision Library. http://ai.stanford.edu/~sgould/svl, 2010.

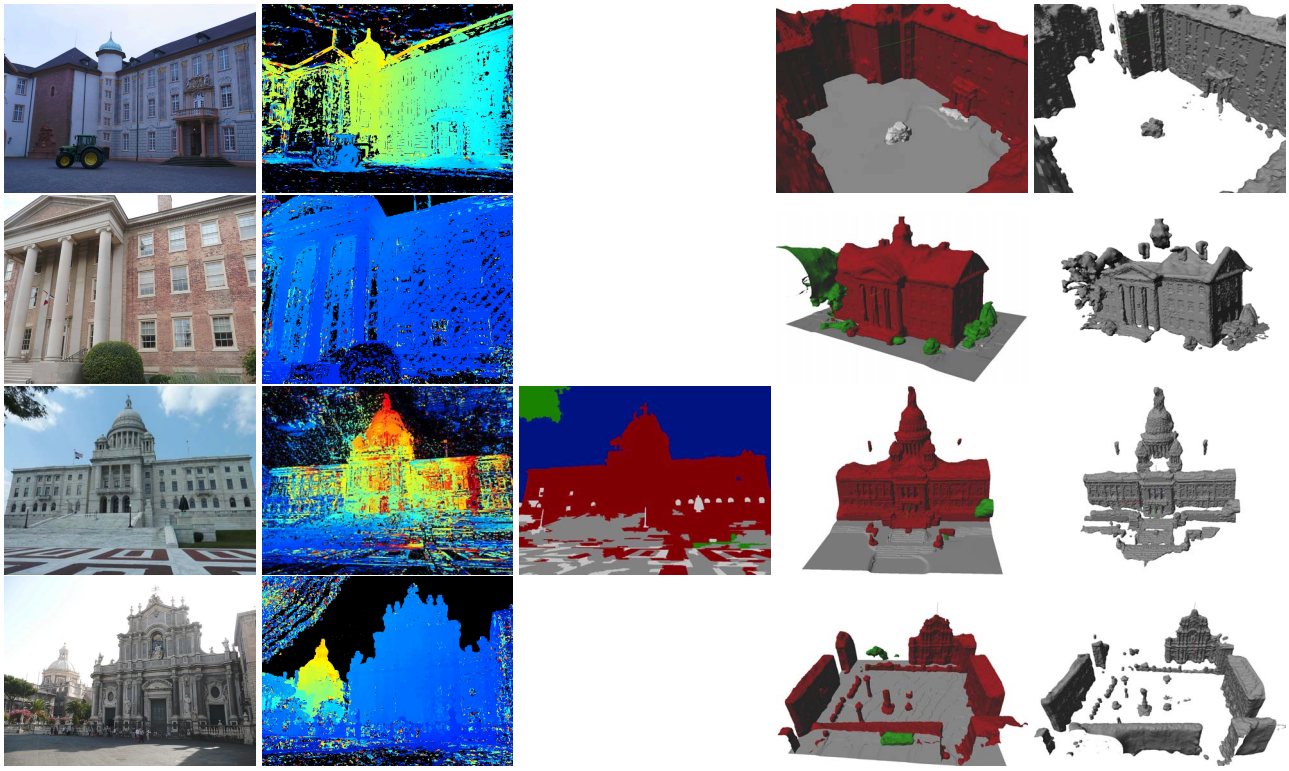

Figure 4: Results for 4 datasets. First row: castle-P30 [21]. Second row to last: Datasets 1 to 3. From left to right: Example input images, example depth map, raw image labeling, our result, tv-flux fusion result; The different class labels are depicted using the following color scheme: building → red, ground → dark gray, vegetation → green, clutter → light gray.

[8] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[9] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proc. CVPR*, 2011.

[10] K. Kolev, T. Pock, and D. Cremers. Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In *Proc. ECCV*, pages 538–551, 2010.

[11] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *Proc. BMVC*, pages 104.1–11, 2010.

[12] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. *SIAM Journal on Imaging Sciences*, 4(4):1049–1096, 2011.

[13] V. Lempitsky and Y. Boykov. Global optimization for shape fitting. In *Proc. CVPR*, 2007.

[14] S. Liu and D. Cooper. A complete statistical inverse ray tracing approach to multi-view stereo. In *Proc. CVPR*, pages 913–920, 2011.

[15] X. Liu, O. Veksler, and J. Samarabandu. Order-preserving moves for graph-cut-based optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1182—-1196, 2010.

[16] J. Melonakos, E. Pichon, S. Angenent, and A. Tannenbaum. Finsler active contours. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):412–423, 2008.

[17] S. Osher and S. Esedoglu. Decomposition of images by the anisotropic Rudin-Osher-Fatemi model. *Comm. Pure Appl. Math.*, 57:1609–1626, 2004.

[18] L. Quan, J. Wang, P. Tan, and L. Yuan. Image-based modeling by joint segmentation. *IJCV*, 75(1):135–150, 2007.

[19] A. Saxena, S. Chung, and A. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 76(1):53–69, 2008.

[20] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, pages 519–526, 2006.

[21] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. CVPR*, 2008.

[22] E. Strekalovskiy and D. Cremers. Generalized ordering constraints for multilabel optimization. In *Proc. ICCV*, 2011.

[23] C. Zach. Fast and high quality fusion of depth maps. In *Proc. 3DPVT*, 2008.

[24] C. Zach, C. Häne, and M. Pollefeys. What is optimized in convex relaxations for multi-label problems: Connecting discrete and continuously-inspired MAP inference. Technical report, MSR Cambridge, 2012.

[25] C. Zach, C. Häne, and M. Pollefeys. What is optimized in tight convex relaxations for multi-label problems? In *Proc. CVPR*, 2012.

[26] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Proc. CVPR*, 2010.

[27] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L^1 range image integration. In *Proc. ICCV*, 2007.