A Patch Prior for Dense 3D Reconstruction in Man-Made Environments

Christian Häne[1] Christopher Zach[2] Bernhard Zeisl[1] Marc Pollefeys[1]

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

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

***Abstract*—Dense 3D reconstruction in man-made environments has to contend with weak and ambiguous observations due to texture-less surfaces which are predominant in such environments. This challenging task calls for strong, domain-specific priors. These are usually modeled via regularization or smoothness assumptions. Generic smoothness priors, e.g. total variation are often not sufficient to produce convincing results. Consequently, we propose a more powerful prior directly modeling the expected local surface-structure, without the need to utilize expensive methods such as higher-order MRFs. Our approach is inspired by patch-based representations used in image processing. In contrast to the over-complete dictionaries used e.g. for sparse representations our patch dictionary is much smaller. The proposed energy can be optimized by utilizing an efficient first-order primal dual algorithm. Our formulation is in particular very natural to model priors on the 3D structure of man-made environments. We demonstrate the applicability of our prior on synthetic data and on real data, where we recover dense, piece-wise planar 3D models using stereo and fusion of multiple depth images.**

I. Introduction

Dense 3D modeling from images often suffers from a lack of strong matching costs, especially in man-made environments. Such environments usually exhibit texture-less surfaces and also non-Lambertian ones violating underlying, e.g. brightness constancy assumptions. The presence of weak data terms must be compensated by strong model priors in order to obtain plausible 3D reconstructions.

In this work we propose to utilize a spatial regularization prior for depth maps inspired by sparse representations. While our energy clearly resembles the dictionary-based energy functionals employed in image processing formulations (e.g. [1]), there are important differences due to the different characteristics of image (intensity) data and depth maps. Most prominently, depth maps representing man-made environments are typically comprised of very few structural elements, but the specific sparsity assumption used for processing intensity images is not necessarily appropriate for depth maps. Thus, the over-complete dictionary used for sparse representations can be substituted by a small one, replacing the sparseness assumption on dictionary coefficients by different priors. One benefit of this reduced model is the significant reduction in the number of unknowns, therefore increasing the efficiency of numerical optimization. Although not being the focus of this work, we anticipate the possibility of extracting the dictionary elements and the statistical prior on the coefficients from training data.

The smoothness prior proposed in this work shares the main motivation with other methods aiming on regularization terms beyond binary potentials. All of these methods are global optimization methods, which are required in order to properly incorporate the spatial smoothness assumptions on the desired solution (i.e. depth map). The most generic (and also the computationally most demanding) approach to handle arbitrary smoothness priors is the formulation using higher-order Markov random fields. [2] proposes to explicitly use second-order regularization priors for stereo in order to obtain smooth surfaces without the typical staircasing effects. Ishikawa and Geiger [3] argue that the human vision system is contrary from current smoothness models; they use second-order priors to penalize large second derivatives of depth and thereby better reflect the behavior of human vision.

Segmentation based stereo approximates pixel-wise higher-order priors by enforcing the constraint that segmented regions of the image should describe smooth 3D surfaces. Often not a general smoothness prior, but a hard constraint is applied, preventing curved surfaces; e.g. [4] uses graph cuts to optimize in the segment domain. Contrary to that is the explicit minimization of a second order prior via alternating, iterative optimization over surface shape and segmentation: Birchfield and Tomasi [5] utilize an affine model to account for slanted surfaces, whereas Lin and Tomasi [6] introduce a spline model to cover more general curved and smooth surfaces. Finally [7] groups pixels of similar appearance onto the same smooth 3D surface (planes or B-splines) and regularizes the number of surfaces, inherently employing soft-segmentation. A parametric formulation for optical flow explicitly aiming on favoring piecewise planar results (instead of generally smooth, or piecewise constant flow fields) is presented in [8]. A recently proposed generalized notion of total varation [9] is also applied to merge several depth maps [10]. This model can be seen as a special case of our formulation.

The contribution of our work is based on the introduction of *patch-based* priors, modeling higher-order priors by

means of a dictionary-based approach. The motivation for this patch based regularization is twofold. First, as dictionaries can be rather small, inference is fast; and second, computational complexity does not increase considerably with the usage of larger patches, allowing for regularization beyond triple cliques and second-order priors. In the following we present the corresponding energy formulation and concentrate on explaining the different choices for the data fidelity term, the dictionary and coefficient priors (Section II). An efficient inference framework is described in Section III to solve for depth maps and dictionary coefficients. Fusion of several noisy depth maps using the patch-based prior is detailed in Section IV, and a computational stereo approach is presented in Section V. Finally, Section VI concludes with a summarizing and prospective discussion of our approach.

II. Proposed Formulation

This section describes the basic energy functional used in our approach and discusses the relation to existing methods. Let Ω be the image domain (usually a 2D rectangular grid), and $\phi_{\mathbf{p}}(\cdot)$ is a family of functions for $\mathbf{p} \in \Omega$ modeling the data fidelity at pixel $\mathbf{p}$. $\phi_{\mathbf{p}}$ is assumed to be convex. Further, let $R_{\mathbf{p}} : \mathbb{R}^{\Omega} \rightarrow \mathbb{R}^{N}$, be a function extracting an image patch centered at $\mathbf{p}$. Since we allow different shapes for the extracted patches (and therefore several different functions $R_{\mathbf{p}}$), we will use another index k to indicate the shape geometry, i.e. $R^k_{\mathbf{p}}$. By allowing different shape geometries for different dictionary sets, the size of the dictionary can be reduced substantially by not using a e.g. square patch uniformly. We utilize the following energy model:

$$E(u,\alpha) = \int \phi_{\mathbf{p}}(u_{\mathbf{p}}) + \eta \sum_k \left\| R^k_{\mathbf{p}} u - \mathcal{D}^k \alpha^k_{\mathbf{p}} \right\| + \psi(\nabla \alpha_{\mathbf{p}})\, d\mathbf{p}, \tag{1}$$

where u is the desired depth (respectively disparity) map, $\alpha^k : \Omega \rightarrow \mathbb{R}^{|\mathcal{D}^k|}$ are the coefficients for the dictionary $\mathcal{D}^k$, η and μ are positive constants, and $\psi(\cdot)$ is a convex function. The individual terms have the following interpretation:

1) The first term, $\phi_{\mathbf{p}}(u_{\mathbf{p}})$, is the data fidelity term at pixel $\mathbf{p}$, and measures the agreement of u with the observed data.
2) The second term, $\sum_k \|R^k_{\mathbf{p}} u - \mathcal{D}^k \alpha^k_{\mathbf{p}}\|$, penalizes deviations of u from a pure dictionary generated patch in a region $R^k_{\mathbf{p}}$ containing $\mathbf{p}$. As distance measure we use either the L_1 norm $\|\cdot\|_1$, or the L_2 (Euclidean) norm, $\|\cdot\|_2$. The choice of $\|\cdot\|_1$ allows u to locally deviate from the predicted patch $\mathcal{D}^k \alpha^k_{\mathbf{p}}$ at a sparse set of pixels, whereas selecting $\|\cdot\|_2$ resembles group Lasso [11] leading to a sparse set of patches in disagreement with $\mathcal{D}^k \alpha^k_{\mathbf{p}}$.
3) The last term, $\psi(\nabla \alpha_{\mathbf{p}})$, adds spatial regularization on the coefficients α. Choosing $\psi(\nabla \alpha_{\mathbf{p}}) = \mu \|\nabla \alpha_{\mathbf{p}}\|$ (for some positive weight μ) leads to piece-wise constant solutions for α.

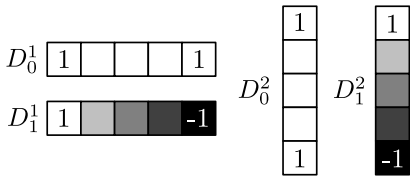


Figure 1. Piecewise planar dictionary elements of length 5.

A. Choice for $\mathcal{D}^k$

The dictionary $\mathcal{D}^k$ determines how the solution should look like locally. It could be acquired by learning it from given training data. By including a sparsity term on the coefficients α^k it would even be possible to use over-complete dictionaries. As this work focuses on the reconstruction of man-made environments we define the dictionary as follows. Man-made environments, and in particular indoor ones, are dominated by locally planar surfaces. Hence it is natural to strongly favor piece-wise planar depth (or disparity) maps on the absence of strong image observations. For the pinhole camera model, it is easy to see that planar surfaces in 3D correspond to locally linear disparity maps (i.e. linear in $1/$depth but not in the depth itself) Consequently, although we use the terms depth map and disparity map equivalently, the smoothness prior is always applied on the disparity representation. For piecewise planar disparity maps the utilized dictionary is very compact and contains only four elements (see Fig. 1),

$$\begin{aligned} D_0^1 = \mathbf{1}^T \quad & D_1^1 = \big(\frac{2i}{P_{\text{length}} - 1} - 1\big)_{i=0}^{P_{\text{length}}-1} \\ D_0^2 = \mathbf{1} \quad & D_1^2 = (D_1^1)^T, \end{aligned} \tag{2}$$

i.e. D_1^1 and D_1^2 are the horizontal and vertical linear gradients, respectively. The coefficients corresponding to the constant elements D_0^1 and D_0^2, α_0^1 and α_0^2, essentially cover the absolute disparity, whereas α_1^1 and α_1^2 represent the local slope in horizontal and vertical image direction. P_{length} denotes the length of the patch.

B. The Choice $\psi(\cdot)$

In order to favor piecewise planar (but not piecewise constant) disparity maps piece-wise constant solutions for the coefficients of the linear gradient elements should be preferred. As we do not want to penalize the actual disparity there is no regularization on the coefficients belonging to the constant dictionary elements. Piece-wise constant solutions can be favored by using the total variation as a penalization:

$$\psi(\nabla \alpha^k_{\mathbf{p}}) = \mu \|\nabla(\alpha^k_{\mathbf{p}})_2\| \tag{3}$$

We either use the isotropic L_2 total variation or the anisotropic L_1 total variation.

C. Choices for $\phi(\cdot)$

The family of functions $\phi_{\mathbf{p}}$ represents the fidelity of the solution u to the observed data at pixel $\mathbf{p}$. We discuss two important choices for $\phi_{\mathbf{p}}$.

In the application of merging multiple depth maps, a varying number of depth measurements (including none) is given for each pixel. Hence, the choice of $\phi_{\mathbf{p}}$ is reflecting the underlying noise model in the input depth maps and penalizes the distance to the given depth measurements $u^l_{\mathbf{p}}$. Depth estimates originating from visual information are often subject to quantization, therefore we use a "capped" L_1-norm to allow deviations within the quantization level. To combine multiple measurements the individual distances are summed up:

$$\phi_{\mathbf{p}}(u_{\mathbf{p}}) = \sum_l \lambda^l_{\mathbf{p}} \max\{0, |u_{\mathbf{p}} - u^l_{\mathbf{p}}| - \delta\}, \tag{4}$$

where $\lambda^l_{\mathbf{p}}$ is a weight that can be specified for each depth measurement. This enables down-weighting of measurements with high matching costs or non-unique matches. This particular choice of $\phi_{\mathbf{p}}$ implicitly fills in missing depth values by setting all weights of missing values to zero.

Another choice for $\phi_{\mathbf{p}}$ addresses the task of computing depth/disparity maps between images. $\phi_{\mathbf{p}}$ may directly measure the similarity of pixel intensities in the reference and matching images, I_0 and I_1, respectively:

$$\phi_{\mathbf{p}}(u_{\mathbf{p}}) = \lambda \big| I_1(u_{\mathbf{p}}) - I_0 \big|, \tag{5}$$

where λ is the weight on the data term. Since with this choice $\phi_{\mathbf{p}}$ is not convex, the definition above is usually replaced by its first order approximation with respect to the linearization point u^0,

$$\phi_{\mathbf{p}}(u_{\mathbf{p}}) = \lambda \big| I_1\big(u^0_{\mathbf{p}}\big) + (u_{\mathbf{p}} - u^0_{\mathbf{p}}) \nabla_u I_1 - I_0 \big|. \tag{6}$$

More generally, any image similarity function can be evaluated locally in the neighborhood of u^0 and its convex surrogate utilized for $\phi_{\mathbf{p}}$ (e.g. the second order approximation of matching costs proposed in [12]). By using only a local approximation of the matching score function, the numerical procedure to find a minimizer u needs to be embedded into a coarse-to-fine framework (or alternatively, an annealing method like [13]).

III. Determining u and α

This section addresses the determination of the unknown disparity map u and corresponding coefficients α for a given dictionary and a coefficient prior. Although the functional in Eq. 1 is convex with respect to the unknowns u and α, it requires minimizing a non-smooth energy. There is a large set of methods for optimizing non-smooth convex problems with additive structure, and many of those contain the proximity operator $\mathrm{prox}_{\sigma f}$,

$$\mathrm{prox}_{\sigma f}(\hat{x}) = \arg\min_x \frac{1}{2\sigma} \|x - \hat{x}\|^2 + f(x) \tag{7}$$

for a convex function f, as a central element. Due to its algorithmic simplicity we chose the primal-dual method proposed in [14], which can be intuitively described as combined gradient descent (in the primal domain) and ascent (in the dual domain) followed by suitable proximity operations. Eq. 1 can be written as convex-concave saddle-point problem (recall that the energy is minimized with respect to u and α),

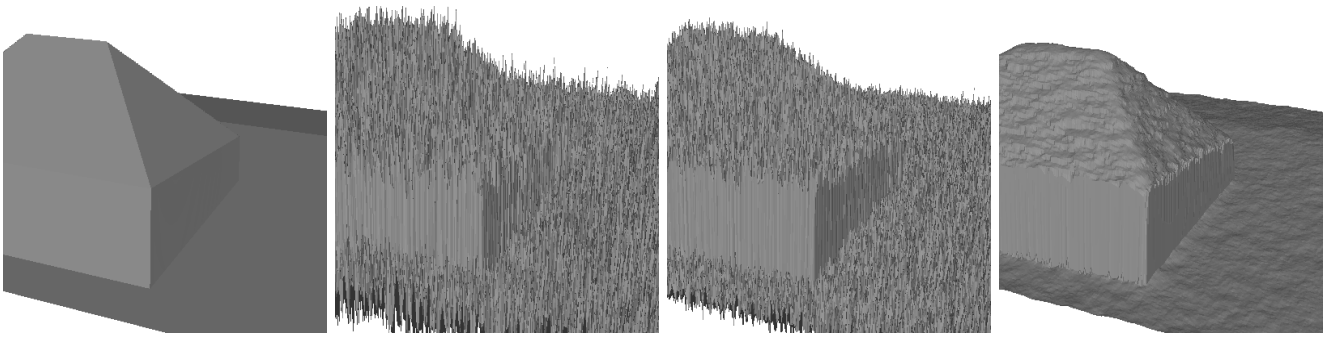
$$\begin{aligned} E(u,\alpha) &= \int \phi_{\mathbf{p}}(u_{\mathbf{p}}) + \eta \sum_k \left\| R^k_{\mathbf{p}} u - \mathcal{D}^k \alpha^k_{\mathbf{p}} \right\|_1 \\ &\quad + \mu \|\nabla \alpha_{\mathbf{p}}\|_2 \, d\mathbf{p} \\ &= \max_{\mathbf{q},\mathbf{r}} \int \phi_{\mathbf{p}}(u_{\mathbf{p}}) + \sum_k \mathbf{q}^T_k \big(R^k_{\mathbf{p}} u - \mathcal{D}^k \alpha^k_{\mathbf{p}} \big) \\ &\quad + \mathbf{r}^T \nabla \alpha_{\mathbf{p}} \, d\mathbf{p} \\ &\quad \text{subject to} \quad \|\mathbf{q}\|_\infty \le \eta, \quad \|\mathbf{r}\|_2 \le \mu. \end{aligned} \tag{8}$$

Here the constraints on $\mathbf{q}$ and $\mathbf{r}$ are formulated for the L_1 norm on the reconstruction error in the primal, and for the isotropic total variation regularizer for α. By choosing different norms the constraints change accordingly. The primal-dual method requires the application of respective proximity operations for the constraints on the dual variables, $\|\mathbf{q}\|_\infty \le \eta$ and $\|\mathbf{r}\|_2 \le \mu$, which are projection steps into the corresponding domain (i.e. element-wise clamping to $[-\eta, \eta]$ and length normalization whenever $\|\mathbf{r}\|_2 > \mu$, respectively).

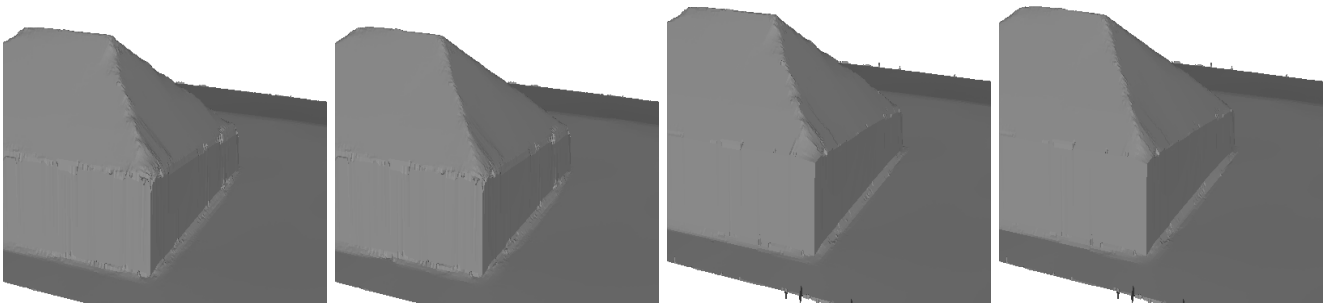
This defines the general framework for using the primal-dual method, and the implications of the different choices of ϕ on the optimization method are discussed in the following section.

IV. Piecewise Planar Depth Map Fusion

Our depth map fusion takes multiple nearby depth maps and combines them in a robust and regularized way to achieve higher quality. One of the input depth maps acts as a reference view, and all other depth images are warped to the reference view-point (using OpenGL-based mesh rendering). Thus, the task is to recover a single depth map $u_{\mathbf{p}}$ from multiple depth measurements $u^l_{\mathbf{p}}$ per pixel. As mentioned in Section II-C we utilize a "capped" L_1 distance accounting for the discrete/quantized nature of depth values. We introduce additional dual variables $\mathbf{s}$, and rewrite the



(a) Ground truth (b) Noisy input (c) Median fusion (d) Huber-TV



(e) TV/L_1+dict./L_2 (f) TV/L_2+dict./L_2 (g) TV/L_1+dict./L_1 (h) TV/L_2+dict./L_1

Figure 2. Top row: synthetic ground truth, one out of 5 noisy inputs, median fusion, Huber-TV fusion. Bottom row: results with the proposed piecewise planar prior with different norms for the dictionary and TV-term. From left to right: TV/L_1 + dictionary/L_2, TV/L_2 + dictionary/L_2, TV/L_1 + dictionary/L_1, TV/L_2 + dictionary/L_1.

primal energy into a saddle-point one and obtain

$$\begin{aligned} E(u,\alpha) = &\int \sum_l \lambda_{\mathbf{p}}^l \max\{0, |u_{\mathbf{p}} - u_{\mathbf{p}}^l| - \delta\} \\ &+ \eta \sum_k \left\| R_{\mathbf{p}}^k u - \mathcal{D}^k \alpha_{\mathbf{p}}^k \right\|_1 + \mu \|\nabla \alpha_{\mathbf{p}}\|_2 \, d\mathbf{p} \\ = &\max_{\mathbf{q},\mathbf{r},\mathbf{s}} \int \sum_l s_{\mathbf{p}}^l (u_{\mathbf{p}} - u_{\mathbf{p}}^l) + |s_{\mathbf{p}}^l| \delta \\ &+ \sum_k \mathbf{q}_k^T \left(R_{\mathbf{p}}^k u - \mathcal{D}^k \alpha_{\mathbf{p}}^k \right) + \mathbf{r}^T \nabla \alpha_{\mathbf{p}} \, d\mathbf{p} \quad (9) \\ &\text{s. t.} \quad \|\mathbf{q}\|_\infty \leq \eta, \quad \|\mathbf{r}\|_2 \leq \mu, \quad |s_{\mathbf{p}}^l| \leq \lambda_{\mathbf{p}}^l. \end{aligned}$$

The proximity operator for the additional constraint is again a projection to the feasible set for $s_{\mathbf{p}}^l$. It is possible to compute the proximity operator for the above data term ϕ directly (without introducing additional dual variables), but this is relatively expensive (see e.g. [15]) and hinders data-parallel implementations.

A. Synthetic Data

We evaluate the behavior of the proposed fusion method for different choices of L_1/L_2-norms in the reconstruction and the TV terms by using a synthetic height-map of a man-made scene (Fig. 2(a)). The input to the fusion is generated by adding zero-mean Gaussian noise with $\sigma = 0.8$ to the ground-truth height-map, which is in the range $[0, 10]$ (see Fig. 2(b) for one of the noisy input depths). Simply taking the pixel-wise median is clearly not sufficient to return a convincing result (Fig. 2(c)), hence enforcing spatial smoothness is required. Adding a smoothness prior via Huber-TV [10] (i.e. enforcing homogeneous regularization for small depth variations and a total variation prior at large depth discontinuities) still results in staircasing artefacts (Fig. 2(d)).

Figs. 2(e-h) depict the results for our proposed method using different combinations of L_1 and L_2 penalizers for the reconstruction error, $\|R_{\mathbf{p}}^k u - \mathcal{D}^k \alpha_{\mathbf{p}}^k\|$, and for the TV regularizer, $\|\nabla \alpha_{\mathbf{p}}\|$. The parameters were chosen as follows: the patch width was fixed to 5, $\lambda_{\mathbf{p}}^l = 1.5$ and $\mu = 10$, to adapt to the different penalization of the reconstruction error, we set $\eta = 1$ for L_1 penalization and $\eta = 1.5$ for L_2 penalization. For the datacost we used $\delta = 0$, which means L_1 distance penalization. In general, all results look rather similar. There are some artifacts at the hip of the roof when using a TV-L_1 penalization for the dictionary coefficients $\alpha_{\mathbf{p}}$. An L_2-norm penalization in the dictionary term slightly cuts-off the edges at the eaves of the roof. The combination of an L_1-norm in the dictionary term and and an isotropic L_2-norm total variation penalization of the dictionary coefficients visually gives the best solutions. In

the remainder of the document we only use this way of penalization.

B. Real-World Data

For our experiments with real-world data we took datasets with 25 images each[1]. To obtain the camera poses we run a publicly available SfM software [16]. Input depth maps are obtained by running plane sweep stereo on 5 input images using a 3×3 ZNCC matching score and best half-sequence selection for occlusion handling. We use semi-global matching [17] to extract depth maps from the ZNCC cost volume, thereby obtaining five depth maps used for subsequent fusion. The depth maps are warped to the reference view by meshing and rendering the depth maps. Implausible triangles corresponding to huge depth discontinuities (i.e. most likely occlusion boundaries) are culled. Further, warped depth values with a corresponding ZNCC matching score smaller than 0.4 are also discarded. For the fusion parameters we used the same settings for all datasets. The patch width was set to 3, $\eta = 1$. $\mu = 5$, $\lambda^l_{\mathbf{p}} = 0.3$ and $\delta = 0.015$. For the Huber-TV fusion we used the same "capped"-L_1 datacost also with $\delta = 0.015$ and the Huber parameter was set to 0.015 as well. In Figs. 3 and 4 we show results for piece-wise planar fusion and Huber-TV fusion on the same input data. Although the Huber-TV fusion aims on reducing the staircasing effect of the standard TV, there are still visually distracting artifacts in the rendered 3D-Model. When utilizing the proposed piece-wise planar structure prior the rendered 3D-Model is visually much more appealing especially when rendered with texture.

Additional results and a video showing the 3D models can be found in the supplementary material.

V. Piece-Wise Planar Depth From Stereo

We can directly incorporate an image matching function (respectively a convex surrogate) for the data fidelity term $\phi_{\mathbf{p}}$ in Eq. 1. We use the L^1 distance between Sobel-filtered image patches as basic matching costs (as suggested in [18], but we utilize only a 3-by-3 window instead of the suggested 9-by-9 one to avoid over-smoothing), and convexify the sampled matching costs using a quadratic approximation as proposed in [12]. The necessary proximal step for this choice of ϕ is given by

$$\text{prox}_{\phi}(u) = \frac{u + \lambda \ddot{c} u^0 - \lambda \dot{c}}{1 + \lambda \ddot{c}}, \tag{10}$$

where u^0 is the current linearization point used to sample the matching cost c, and $\dot{c}$ and $\ddot{c}$ are the first- and second-order derivatives of c with respect to disparity changes, computed via finite differences. Since the true matching cost approximation is only valid in a neighborhood of u^0, we explicitly add the box constraint $|u - u^0| \leq 1$ to limit the range of disparity updates to 1 pixel. Adding this constraint to ϕ means that the proximal operator above is followed by a clamping step to force u to be in $[u^0 - 1, u^0 + 1]$.

Further, the numerical procedure has to be embedded into a coarse-to-fine framework, with optional multiple cost sampling (i.e. image warping) steps per pyramid level. Without a suitable initialization this comes at the risk of missing small structures in the final disparity map (which is an intrinsic problem of all multi-scale methods).

Figs. 5 and 6 illustrate the difference between stereo with a total variation smoothness prior and the piecewise planar prior using the proposed formulation. The weighting between the data fidelity and the smoothness term in the TV model is selected, such that the results of both formulations are visually similar. As expected, using the TV regularizer leads to visible staircasing, in particular in textureless regions, which can be overcome by using the patch-based prior. Using 8 CPU cores the stereo approach generates usable depth maps for 384×288 images at 5 Hz, and more than 7 Hz can be achieved for 512×384 images using a GPU implementation (measured on an NVidia GTX 295). Consequently, it is conceivable e.g. to enable live and dense reconstruction for challenging indoor environments in the spirit of [19], [20].

VI. Conclusion and Future Work

In this work we present an energy formulation for depth map recovery utilizing a patch-based prior, and apply the proposed framework to depth map fusion and computational stereo. We describe an efficient method to infer depth from given image data. One major difference of our approach to second order regularization like [8], [2] is, that our formulation is able to consider much larger neighborhoods, without the computational drawbacks of higher-order MRFs with large clique sizes.

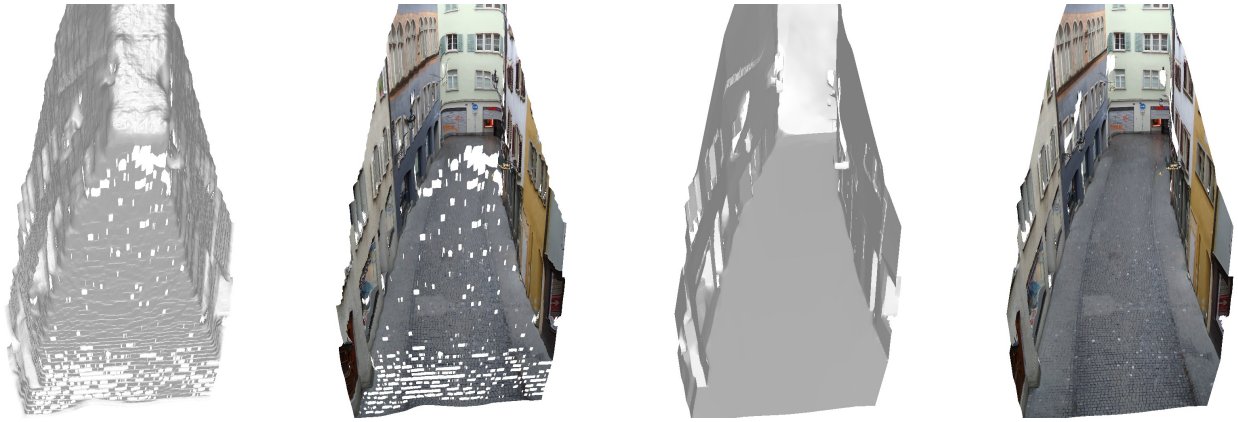
The focus of this work was on the model formulation and on the inference step to obtain depth maps, with the assumption that the patch dictionary and the priors on dictionary coefficients are known. Learning patch elements and coefficient priors from training data is subject of future work. Beside replacing the manual design of dictionaries by an automated training phase, linking dictionary elements with appearance based classifiers for category detection is an interesting future topic. Joint optimization for depth and semantic segmentation is recently addressed in [21], where absolute depth and object categories are directly combined. Linking the local depth structure (i.e. the dictionary coefficients) with object segmentation seems to be a more powerful approach.

Acknowledgment

This work was supported in parts by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant n.269916 (V-Charge).

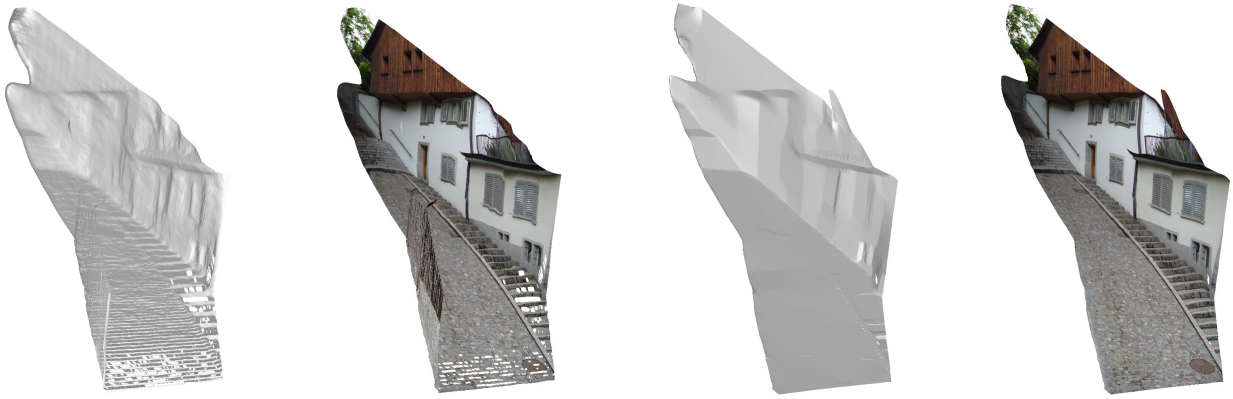
[1]Datasets are available at `http://people.inf.ethz.ch/chaene/`

Input image | Input depth map | Huber-TV fusion | Piecewise planar fusion



Huber-TV fusion | Piecewise planar fusion

Input image | Input depth map | Huber-TV fusion | Piecewise planar fusion



Huber-TV fusion | Piecewise planar fusion

Figure 3. Outdoor depth map fusion results. Odd rows: One out of 25 input images, one out of 5 generated input depth maps, depth map form Huber-TV fusion, depth map with proposed piece-wise planar prior. Even rows: Untextured and textured 3D-Model. Left from Huber-TV fusion and right with piecewise planar prior.

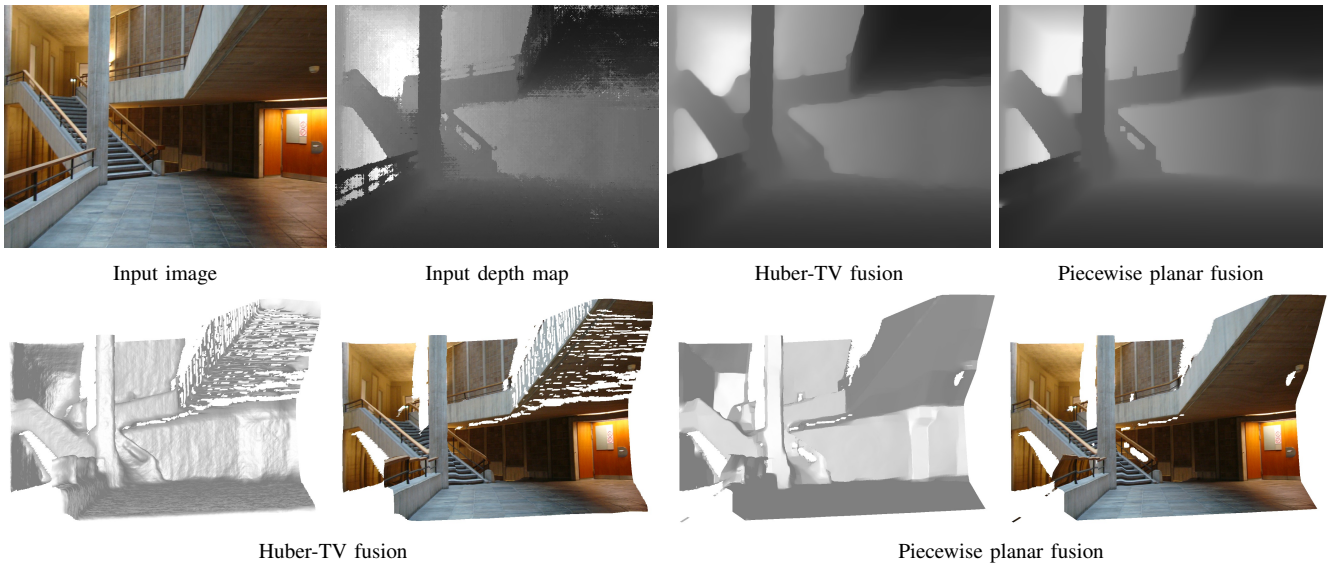


Figure 4. Indoor depth map fusion results. First row: One out of 25 input images, one out of 5 generated input depth maps, depth map form Huber-TV fusion, depth map with proposed piece-wise planar prior. Second row: Untextured and textured 3D-Model. Left from Huber-TV fusion and right with piecewise planar prior.

Figure 5. From left to right: left and right input image, stereo result with TV prior, stereo result with piece-wise planar prior

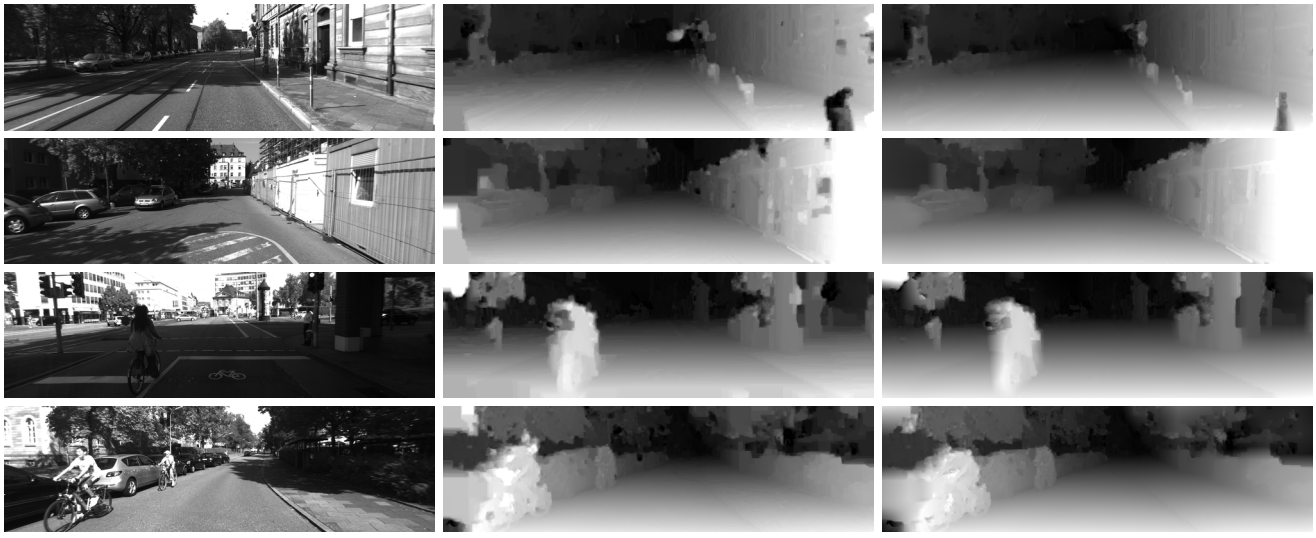


Figure 6. Results for the four urban data sets (available from http://rainsoft.de/software/libelas.html): left input image (672x196), depth from stereo using the TV prior, depth from stereo using the piecewise planar prior.

REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736 –3745, 2006.

[2] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, "Global stereo reconstruction under second-order smoothness priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, pp. 2115–2128, 2009.

[3] H. Ishikawa and D. Geiger, "Rethinking the prior model for stereo," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 526–537.

[4] L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2004.

[5] S. Birchfield and C. Tomasi, "Multiway cut for stereo and motion with slanted surfaces," in *IEEE International Conference on Computer Vision (ICCV)*, 1999, p. 489.

[6] M. Lin and C. Tomasi, "Surfaces with occlusions from layered stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1073–1078, 2004.

[7] M. Bleyer, C. Rother, and P. Kohli, "Surface stereo with soft segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[8] W. Trobin, T. Pock, D. Cremers, and H. Bischof, "An unbiased second-order prior for high-accuracy motion estimation," in *Proc. DAGM*, 2008.

[9] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM J. Imaging Sci.*, vol. 3, no. 3, pp. 492–526, 2010.

[10] T. Pock, L. Zebedin, and H. Bischof, "Rainbow of computer science," 2011, ch. TGV-Fusion, pp. 245–258.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of The Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[12] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[13] F. Steinbrücker, T. Pock, and D. Cremers, "Large displacement optical flow computation without warping," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[14] A. Chambolle and T. Pock, "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging," *Journal of Mathematical Imaging and Vision*, pp. 1–26, 2010.

[15] C. Zach, T. Pock, and H. Bischof, "A globally optimal algorithm for robust TV-L^1 range image integration," in *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[16] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1426–1433.

[17] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 807–814.

[18] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference on Computer Vision*, 2010.

[19] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[20] J. Stuehmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition (Proc. DAGM)*, 2010, pp. 11–20.

[21] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr, "Joint optimization for object class segmentation and dense stereo reconstruction," *Int. Journal of Computer Vision*, pp. 1–12, 2011.