Department of Computer Science
George Mason University
Technical Report Series

4400 University Drive MS#4A5
Fairfax, VA 22030-4444 USA
http://cs.gmu.edu/
703-993-1530

Multi-view Superpixel Stereo in Man-made Environments

Branislav Mičušík
bmicusik@gmu.edu

Jana Košecká
kosecka@cs.gmu.edu

Technical Report GMU-CS-TR-2008-1

Abstract

Man-made environments possess many regularities which can be efficiently exploited for 3D dense reconstruction from multiple widely separated views. We present an approach utilizing properties of piecewise planarity and restricted number of plane orientations to suppress the ambiguities causing failures of standard dense stereo methods. We formulate the problem of the 3D reconstruction in MRF framework built on an image presegmented into superpixels. Using this representation, we propose novel robust cost measures, which overcome many difficulties of standard pixel-based formulations and handles favorably problematic scenarios containing many repetitive structures and no or low textured regions. We demonstrate our approach on several low textured, wide-baseline scenes demonstrating superior performance compared to previously proposed methods.

1 Introduction

Previous approaches to acquisition of 3D dense models from multiple views differ in the type of chosen geometric primitives, estimation algorithms as well as level of human interaction. There exist several systems for completely automated recovery of camera motion and 3D structure of the scene, e.g. [17, 1]. In many instances the general methods lack for robustness, demonstrated in Fig. 1 and are well conditioned only in restricted scenarios. Alternatively, systems that have enjoyed success in limited domains typically employ structural information and require some level of interaction [3, 13].

With the increased interest in modeling of urban environments as well as richness of the various geometric configurations of basic building blocks, we want to extend the class of 3D dense reconstruction techniques which can be used in automated or semi-automated manner in cases where wide-baseline views are available. The typical problems causing the failure of standard methods are lack of textured areas, presence of repetitive textures, or large changes of image properties across views. In this paper we cope with those problems by taking into account unique properties of the man-made environments like piecewise planarities and dominant plane orientations. We propose how to encode these properties into a 3D dense reconstruction pipeline and show that

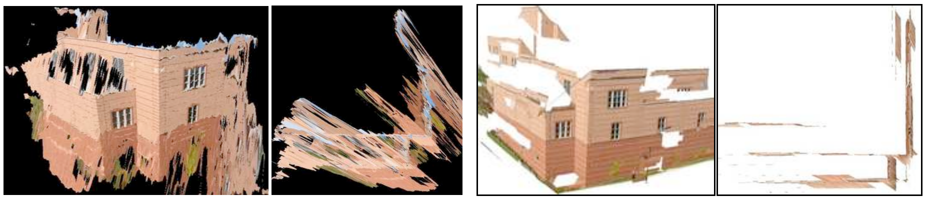

Figure 1: Comparison of the 3D models created by the ARC 3D webservice [17] employing state-of-the-art techniques (two most left images) and by our proposed method. Right hand side image from each group shows top view on the model. Notice more consistent and complete model created by our method. Input images are shown in Fig. 3.

they yield more accurate and visually pleasing results, shown for instance in Fig. 1. This work complements recent work in multi-view stereo which typically focuses on often highly textured scenes [16].

The main contribution of this work is the choice of image representation by superpixels, opposed to standard used single pixels, and novel energy term formulations harvesting advantageous properties of such representation in man-made environments. We employ the sweeping strategy encapsulated in a Markov Random Field (MRF) framework similarly to [6, 5, 18], described in more detail in Sec. 2.2. We rely only on the fact that the 3D patches projected into superpixels are planar which enables us to reliably recover many planar structures. Such assumption allows more freedom than the predefined geometric primitives like windows, doors etc., sometimes used in 3D reconstruction pipelines [20]. Our proposed superpixel based formulation is beneficial for the man-made environments in many aspects:

1. It solves the ambiguities present in standard dense stereo methods at places with no or low texture. Those places are merged in superpixels and are treated as larger entities implicitly restricting their possible 3D positions.

2. It is more robust to a camera misalignment enabling us to handle favorably wide-baseline settings with illumination or camera exposure changes across views. Single pixel formulations usually reliably work on dense narrow-baseline sequences only.

3. More robust photoconsistency measures can be designed over superpixels covering larger areas compared to standard small square windows centered at pixels.

4. Significant reduction of computational complexity can be achieved as the number of nodes in the graph built on superpixels is much smaller, by a factor of 1000, compared to graphs built on all image pixels.

The superpixels are widely used in image segmentation methods usually followed by a recognition stage. It has recently been shown in [15] how the superpixels can be utilized in connection to the 3D stereo reconstruction using single image 3D reconstructions based on pre-learned priors from laser scanners.

The structure of the paper is the following. In Sec. 2 we first formulate the problem being solved here, second we compare previous and our suggested solution, and

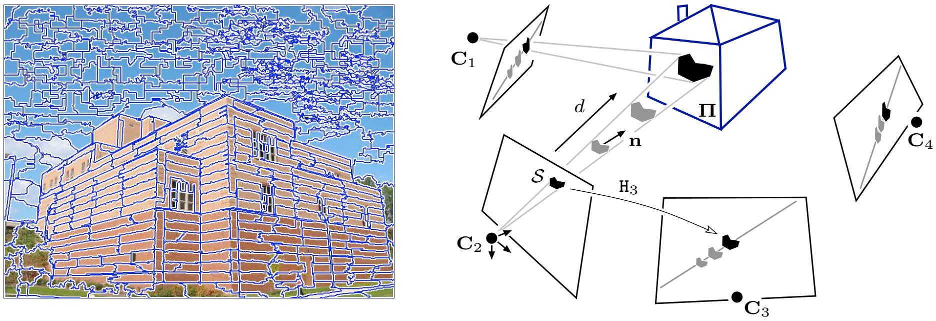

Figure 2: *Left:* Superpixels being swept detected by [4] on a reference image from the BUILDING dataset. *Right:* The sweeping concept discussed in the paper.

in Sec. 3 we explain our proposed algorithm into detail. Experiments in Sec. 4 show some results supporting the feasibility of the method.

2 Multi-view superpixel stereo

Let us assume a rigid scene of man-made environment observed by widely separated calibrated cameras with known camera projection matrices and a presegmented reference image into superpixels. The number of cameras is arbitrary but more than two and the input images are not required to be rectified. As an output we want to assign a normal and depth in 3D to each superpixel.

Setup: The camera projection matrices are assumed to be of the form $\mathtt{P}_k = \mathtt{K}_k[\mathtt{R}_k\ \mathbf{t}_k]$, where $\mathtt{K}_k$ stands for the camera intrinsic calibration matrix, and $\mathtt{R}_k$, $\mathbf{t}_k = -\mathtt{R}_k\,\mathbf{C}_k$ for the rotation matrix and translation vector of the k-th camera with the center $\mathbf{C}_k$ wrt a reference camera coordinate system [7].

Superpixels: The superpixels partition the image into locally homogeneous irregular patches, usefully containing pixels of same color or forming locally same texture. An example of the superpixel partitioning can be seen in Fig. 2. Let us assume that the superpixels come from projection of 3D *planar* patches. This is a reasonable assumption for the man-made environments our method is designed for. Suitable superpixels are those which respect boundaries and correspond to object boundaries or depth discontinuities in 3D [14, 4].

2.1 Problem statement

The planar assumption of the 3D patches states that a patch we want to reconstruct from its projection into a superpixel must lie at the intersection of the projective cone passing the superpixel boundary and a plane containing the 3D patch, see Fig. 2. Denote such a plane $\mathbf{\Pi} = [\mathbf{n}^\top\ d]^\top$, where $\mathbf{n}$ is the unit normal vector expressed in the reference coordinate system and d is the distance along the normal of that plane to the origin.

All pixels $\mathbf{x}_{ref}$ inside a superpixel $\mathcal{S}$ in the reference image, if being a projection of a planar patch, are mapped through a plane induced homography $\mathtt{H}_k \in \Re^{3\times3}$ into the k-the camera view as $\mathbf{x}_k \simeq \mathtt{H}_k\, \mathbf{x}_{ref}$, where

$$\mathtt{H}_k(\mathbf{\Pi}, \mathtt{P}_k, \mathtt{K}_{ref}) = \mathtt{K}_k \left(\mathtt{R}_k - \mathbf{t}_k\, \mathbf{n}^\top / d\right) \mathtt{K}_{ref}^{-1}, \qquad (1)$$

see [7] for more detail. The only *unknowns* are the plane parameters $\mathbf{n}$, d. As pointed out in [6] one can estimate the unknown parameters by sweeping the plane $\mathbf{\Pi}$ by d along different normals searching for such values $\mathbf{n}^*$ and d^* that map the superpixel $\mathcal{S}$ via homography $\mathtt{H}_k$ at correct positions in other views.

The problem being considered here is to estimate parameters of all superpixel planes $\mathbf{\Pi}_s^*$ simultaneously from initial candidate hypotheses obtained by the sweeping. The goal is to find a solution giving minimal photoconsistency error of the superpixels and their projections in all views while respecting geometric properties of the superpixels and considering smooth changes of the depth and normals of the neighboring superpixels. We want to find $\mathcal{P}^* = \{\mathbf{\Pi}_s^* : s = 1 \dots S\}$ as a Maximum Posterior Probability (MAP) assignment of a MRF, whose graph structure is induced by neighborhood relationships between superpixels. More formally we seek such $\mathcal{P}^*$ that

$$\operatorname*{argmin}_{\mathcal{P}} \left[\sum_s E_{photo} + \lambda_1 \sum_s E_{geom} + \lambda_2 \sum_{\{s,s'\}} E_{norm} + \lambda_3 \sum_{\{s,s'\}} E_{depth} \right], \qquad (2)$$

where E's stand for energies, discussed in detail later, λ's are their weights, $\{s, s'\}$ are neighboring superpixels, and $\mathcal{P}$ is a set of all possible planes for all S superpixels, i.e. $\mathcal{P} = \{\mathbf{\Pi}_s : s = 1 \dots S\}$ and $\mathbf{\Pi}_s = [\mathbf{n}_s^\top \; d_s]^\top$.

2.2 Related work

There are several recent approaches how to cope with the NP-hard problem stated in Eq. (2) and make it computationally tractable.

If the problem is formulated on pixels and photoconsistency term is computed over small windows centered at the pixels [18] there is no need to estimate the normals. The large number of plane distances (depths) is reduced by selecting those depths which correspond to local minima along sweeping direction. The problem is then formulated as the binary MRF solvable by the min-cut solver.

In [5] the problem is solved in an iterative framework. First, they match corners and blob features and estimate their initial depths and normals from triangulation. Second, they pass the depths and normals into a photoconsistency minimization framework based on normalized cross correlation on patches over all views. They filter outliers enforcing visibility consistency and repeat the same procedure a couple of times while adding new neighbors to existing patches at each stage and regularizing over neighbors.

Operating still on pixels but increasing window size for photoconsistency term computation is investigated in [6]. However, larger windows require that normals have to be taken into account as well. Their algorithm is suited for the man-made environments and therefore they consider only three dominant directions of normals. One depth is chosen for each of three normals as a minimum of photoconsistency error along the sweeping ray across all views resulting in a discrete MRF with 3 labels for each pixel.

Although the methods mentioned above operate on pixels, for computing the photoconsistency term, they use windows to make a solution more robust to noise. The windows centered at each pixel are assumed to be a projection of 3D planar patches. This assumption is not valid for many pixels in the man-made environments. Especially for the pixels near boundary of two planes with different normals as the windows start capturing pixels from another plane. Using larger windows yields more stable results, however, there is higher chance that the planar assumption is violated. As a remedy, superpixels naturally follow the boundaries and offer a natural way to avoid those problematic cases, while still maintaining the robustness by operating on many pixels. The photoconsistency term is computed only on pixels inside the superpixels where it is more likely that the pixels belong to the same plane. Moreover, the methods mentioned above typically work in narrow-baseline settings with many views to suppress sensitivity to camera pose inaccuracies.

3 Proposed solution

We restrict the space of directions of normals to dominant directions in the scene, similarly to [6]. The set of depths for each superpixel is restricted to N_{best} local minima along the sweeping direction as opposed to [6] where only one minimum is considered. This is especially important in scenes with low texture and many repetitive patterns. Compared to [6] we operate on more labels and reserve one extra "don't know" label to handle superpixels which are highly ambiguous. Next we describe a novel photometric and geometric consistency criteria used in the energy terms and accommodate the smoothness terms accommodating beneficial properties of the superpixels.

We formulate the problem in Eq. (2) as a discrete *labeling* problem on a graph with fixed number of L labels per superpixel. The labels correspond to possible candidates for depth and normal obtained in the sweeping stage. In the MRF graph structure the superpixels s stand for graph vertices and pairwise connections $\{s, s'\}$ are established between all neighboring superpixels. The costs of labels l_s and pairwise label edges $\{l_s, l_{s'}\}$ are set accordingly to energy terms defined in Sec. 3.2. More details about solving Eq. (2) are given in Sec. 3.3.

In all presented results here, we obtain the dominant directions from detected vanishing points $\mathbf{v}_i$ in the reference image and setting them to $\mathbf{n}_i = \mathtt{K}_{ref}^{-1}\mathbf{v}_i$. In the Manhattan world there are three dominant mutually orthogonal directions and there are many methods to detect them automatically [10, 8]. Alternatively, the normal directions can be obtained from sparse cloud of reconstructed 3D points as proposed in [6]. However, any directions can be added without further changes of the formulation.

3.1 Sweeping stage

The sweeping procedure is done for each superpixel $\mathcal{S}_s$ independently. Therefore we drop the subscript s in next algorithm for better clarity unless necessary and explain the procedure on one superpixel only. The whole process is repeated for all superpixels and for all considered normals $\mathbf{n}_i$. As an output a matrix $\mathtt{T}^s$ is returned for each superpixel.

1. Sweep a plane $\mathbf{\Pi} = [\mathbf{n}_i^\top \; d]^\top$ in 3D by changing d about Δd along $\mathbf{n}_i$ from d_{min} to d_{max} given as parameters. Repeat the steps 2-4 for all those d.

2. Obtain projections of a superpixel s in all views: Compute the homography matrix from Eq. (1). Map the points from a rim of the considered superpixel s into the other views, create polygon approximation of the boundary and select all interior points, see first row images in Fig. 3.

3. Photometrically normalize each superpixel projection: Compute the chromacity vector $\mathbf{c}_k = \frac{[\mu_k^R \; \mu_k^G]^\top}{\mu_k^R + \mu_k^G + \mu_k^B}$, where μ_k^R, μ_k^G, and μ_k^B are mean colors of all pixels inside the projection. Transform colors of those pixels such that their mean and variance in each color channel become 0 and 1.

4. Evaluate cost of each superpixel projection: Compute histogram for each color channel from the photometrically normalized pixels of the superpixel projection using Parzen windows with linear kernel and stack them in one vector $\mathbf{h}_k$. We use 20 bins per color channel. Compute histogram difference[1] $\chi_k^2(\mathbf{h}_k, \mathbf{h}_{ref})$. Find the set $\mathcal{K}$ of those views k where whole superpixel projection lies in the image and $\chi_k^2 < \Theta_1$, indicating non-occluded 3D patch for the k-th camera. Evaluate the following cost measure

$$C(d) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} C_k(d) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left(\chi_k^2 + \alpha \|\mathbf{c}_k - \mathbf{c}_{ref}\|^2 \right), \tag{3}$$

   where the second term of $C_k(d)$ encodes chromacity difference of the superpixel and its k-th projection.

5. Find depth candidates: Search for N_{best} minima of $C(d)$ over d as possible candidates for depths with normal $\mathbf{n}_i$, see Fig. 3. Store the candidates represented by $[d \; i \; C(d)]^\top$ by adding them into the matrix $\mathtt{T}^s$ as new columns.

We investigated other possibilities of $C(d)$ in Eq. (3), as *i)* sum of all $C_k(d)$ [6] and *ii)* as a function created by summing Gaussians put at local extrema of $C_k(d)$ [18]. However, the mean of inliers in Eq. (3) provided the most stable results in the sense of localizing possible depth candidates and handling occlusions. In all our experiments we used $\Theta_1 = 2$ and $\alpha = 10$.

The photometric normalization in step 3 is essential for handling scene illumination and camera exposure changes, which are often present in wide-baseline settings. The idea of the photometric normalization was also favorably used in context of wide-baseline matching in [11]. In case of the pixel based techniques the illumination invariance is achieved only partially by using normalized cross correlation or by hardware solution using directly a gain read out from the camera [6]. The superpixels allow to handle this invariance more robustly. Third image in Fig. 3 shows that even though it was shot with different, automatically adjusted exposure, the histogram still indicates a region with the same properties.

[1] Chi-squared histogram distance is defined as $\chi^2(\mathbf{h}, \mathbf{g}) = \frac{1}{2} \sum_j^{N_{bins}} \frac{(h(j) - g(j))^2}{h(j) + g(j)}$.

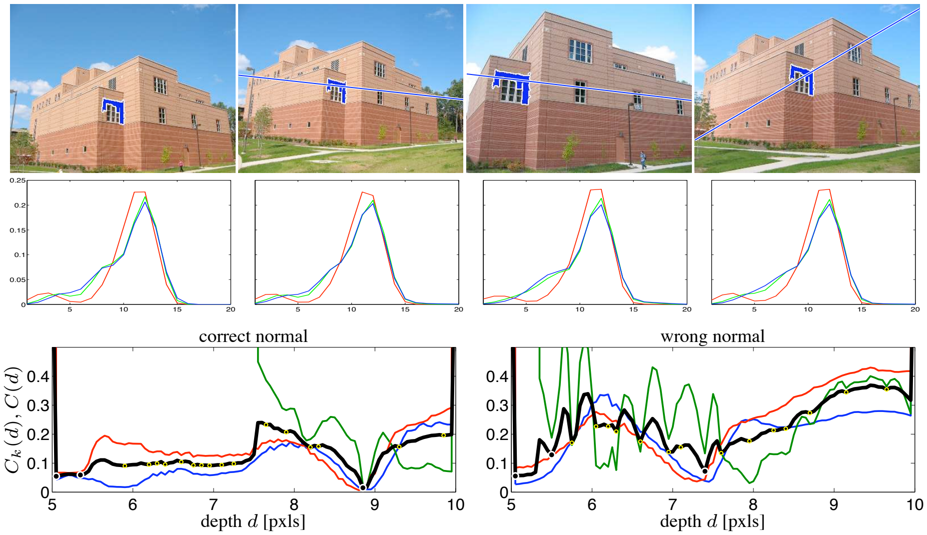

Figure 3: Photoconsistency measure. *First row:* Input images from the BUILDING dataset with an investigated superpixel with its correct projections in all views and epipolar lines for a center of gravity of that superpixel. *Second row:* Histograms in R, G, B color channels computed from pixels of the shown superpixels. *Third row:* Cost measure $C_k(d)$ for each view and one composed cost measure $C(d)$ from Eq. (3) shown in bold black for correct and wrong normal (corresponding to perpendicular walls). All local minima are plotted as small bullets while $N_{best} = 3$ minima are enlarged. Red, Green, Blue colors correspond to fourth, third, and second view respectively. Notice more jagged graph and higher minima for the wrong normal.

3.2 Energy terms

Photoconsistency term. This term penalizes appearance inconsistencies of the superpixels across all views and is set as following

$$E_{photo}(s, l_s) = \mathtt{T}^s_{(3,l_s)}, \tag{4}$$

where l_s is an assigned label for a superpixel s and the number $\mathtt{T}^s_{(3,l_s)}$ is a 3rd row and l_s column in the matrix $\mathtt{T}^s$, built in Sec. 3.1, corresponding to the photometric cost $C(d)$ of a depth candidate for a particular superpixel s.

Geometric term. We employ the assumption that in man-made environments superpixel boundaries are usually aligned with the vanishing directions. The geometric consistency of a superpixel with a particular plane normal is expressed via a deviation of gradient orientation of the pixels along the boundary of the superpixel to two vanishing points corresponding to that plane.

In all our experiments we employ 5-component gradient mixture model for the Manhattan world described in [2]. For each image pixel, the model provides the probability of the pixel lying on an edge, probability to pointing to each of the three vanishing points, see Fig. 4, and the probability of being noise. We take into account only those

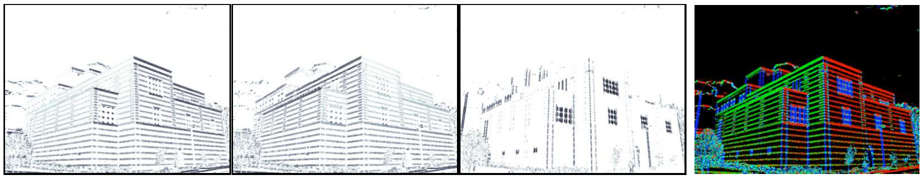

Figure 4: Gradient mixture model [2]. From left: three gradient probability images of being aligned to each of three vanishing points a color-coded membership image of each pixel to one of three vanishing points (red, green, blue), not to be consistent with any (cyan), and to be noise (black).

pixels having the probability of being on an edge higher than being noise. For each of those points a maximum over last 4 probabilities is chosen as a membership of the point to either being consistent with one of the 3 vanishing points or not to be consistent with any, see the most right image in Fig. 4. Then, for a particular superpixel s, we compute a normalized histogram $\mathbf{g}_s(y)$ with four bins $y = \{1, 2, 3, 4\}$ from memberships of all pixels lying along the superpixel boundary.

A probability $p_s(i)$ of the superpixel s to be compatible with a plane with the normal $\mathbf{n}_i$ is captured by $p_s(i) = \sum_{\substack{j=1 \\ j \neq i}}^{3} \mathbf{g}_s(j)$ if $i = \{1, 2, 3\}$ and $p_s(i) = \mathbf{g}_s(4)$ otherwise. The geometric term is then set as follows

$$E_{geom}(s, l_s) = 1 - p_s(\mathtt{T}^s_{(2,l_s)}), \tag{5}$$

where $\mathtt{T}^s_{(2,l_s)}$ indicates which normal is being considered. The term converges to 0 for rectilinear shaped superpixels with boundaries aligned with two vanishing points perpendicular to a considered normal. In a non-Manhattan world with more than 3 non-orthogonal dominant directions in scene, a method with more components handling more vanishing points can be formulated analogously.

Normal term. The normal term is one of the smoothness terms operating on superpixel pairs. This term penalizes the changes of normal directions of neighboring superpixels s and s', and is defined as

$$E_{norm}(s, s', l_s, l_{s'}) = \delta\big(\mathtt{T}^s_{(1,l_s)} \neq \mathtt{T}^{s'}_{(1,l_{s'})}\big), \tag{6}$$

where $\delta(.)$ is 1 when argument is true and 0 otherwise.

Depth term. This smoothness term penalizes the changes in 3D depth of points of the common boundary of two neighboring superpixels. Denote $\mathbf{x} \in \mathcal{S}_s \cap \mathcal{S}_{s'}$ the points on the boundary and $\mathbf{X}(\mathbf{x}, \mathbf{n}, d) = d\, \mathtt{K}_{ref}^{-1}\mathbf{x} / \big(\mathbf{x}^\top \mathtt{K}_{ref}^{-\top}\, \mathbf{n}\big)$ to be a reconstructed inhomogeneous 3D point as an intersection of a sweeping plane and the projective ray $\mathtt{K}_{ref}^{-1}\mathbf{x}$. Then, the smoothness term

$$E_{depth}(s, s', l_s, l_{s'}) = \min\Big(\underset{\mathbf{x} \in \mathcal{S}_s \cap \mathcal{S}_{s'}}{\operatorname{med}} \frac{\|\mathbf{X}(\mathbf{x}, \mathbf{n}_s, d_s) - \mathbf{X}(\mathbf{x}, \mathbf{n}_{s'}, d_{s'})\|}{\|\mathbf{X}(\mathbf{x}, \mathbf{n}_s, d_s)\|}, \Theta_2 \Big), \tag{7}$$

is defined as median of the relative error, where $\mathbf{n}_s = \mathbf{n}_{\mathtt{T}^s_{(2,l_s)}}$, $d_s = \mathtt{T}^s_{(1,l_s)}$ is the considered planar hypothesis and analogously for $\mathbf{n}_{s'}$ and $d_{s'}$. The parameter $\Theta_2 = 0.5$ controls saturation of the measure to avoid strong penalization of true depth discontinuities in the surface.

3.3 MAP solution of MRF

Given a discrete set of planar hypotheses for each superpixel obtained in the sweeping stage, stored as vectors in the matrices $\mathtt{T}_s$, the overall 3D reconstruction can then be formulated as a discrete labeling problem where labels l_s correspond to columns in matrices $\mathtt{T}_s$. In particular, given L labels corresponding to number of considered normals $\times\ N_{best}$ $+1$ (we reserve one label for the "don't know" label), we seek a set of planes $\mathcal{P}$, which minimizes the energy in Eq. (2). The symbol $\mathcal{P}$ in Eq. (2) stands then for the S dimensional vector l which assigns each superpixel s one label l_s representing the best planar hypothesis for that superpixel. The photoconsistency term for the "don't know" labels are set as a constant set to mean $\mathtt{T}^s_{(3,:)}$ over all s and the geometric term is set to 1. The pairwise edges with those labels have a fixed uniform prior.

Recently, very efficient and fast algorithms for solving this type of labeling problem through linear programming relaxation and its Lagrangian dual have been reviewed in [9, 19]. Although finding a global optimum of Eq. (2) is not guaranteed, as the problem is NP-hard, it has been shown that often the optimal solution or one very close to it can be reliably achieved.

3.4 Second iteration utilizing plane priors

After the first run of the labeling algorithm we obtain one depth and normal per superpixel. We propose to re-compute the graph weights on edges utilizing the current estimate and re-run the algorithm. The reason is that the photoconsistency measure in Eq. (3) can miss some correct depths in the first run.

We utilize prior plane probabilities $p(\mathbf{\Pi}_s)$ set as normalized histograms computed from estimated depths of all pixels for each normal separately, see middle column in Fig. 5. In [6] the plane priors are directly computed from triangulated point correspondences which are not considered to be available here. It has been shown there that by incorporating the plane prior $p(\mathbf{\Pi}_s)$ in the Bayesian formulation, assuming the likelihood of the form $p(C(d)\,|\,\mathbf{\Pi}_s) = e^{-\frac{C(d)}{2\sigma^2}}$, and by minimizing negative log-likelihood $\log\big(p(C(d)\,|\,\mathbf{\Pi}_s)\,p(\mathbf{\Pi}_s)\big)$ one ends up with new cost function

$$C_2(d) = C(d) - 2\sigma^2 \log p(\mathbf{\Pi}_s). \qquad (8)$$

We use this cost function as a replacement for the cost function in Eq. (3) in the sweeping algorithm in Sec. 3.1. The second iteration stage with $\sigma = 0.1$ in Eq. (8) contributes to smoother surfaces and help to solve for inconsistencies in the depth estimates remained after the first step, as shown in Fig. 5.

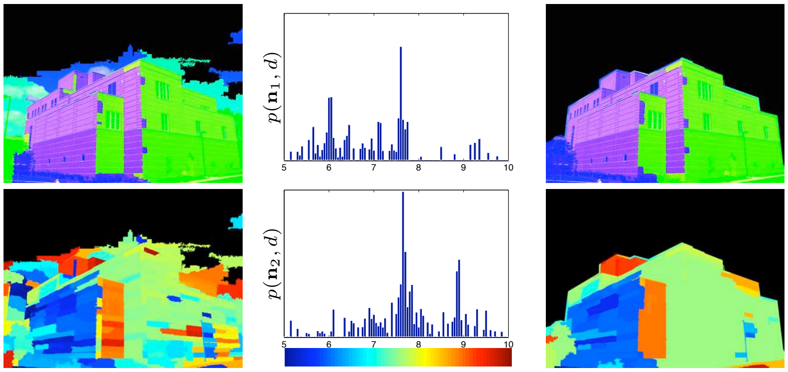

Figure 5: Two stage estimation process. *First column:* Result after the first stage. The top image shows estimated directions of normals (pixels assigned to $\mathbf{n}_1$ have saturated blue, $\mathbf{n}_2$ green channel), the bottom image shows estimated depth encoded in color range. Black color stands for the "don't know" label. *Second column:* Histograms from the estimated depths representing $p(\mathbf{\Pi}) = p(\mathbf{n}, d)$ used as priors for the next stage. Notice two dominant planes in each normal direction. *Third column:* The smoother and consistent result after the second refinement stage. Sky is manually masked out.

4 Experiments

The presented method is demonstrated on several scenes of man-made environments. Videos of 3D reconstructions of all considered sequences can be seen online[2].

The first BUILDING dataset, taken by us, shown in Fig. 3, consists of 4 wide-baseline images containing repetitive texture and large illumination changes. Camera poses were obtained by [17]. In this experiment only two normals are used since the ground plane is not visible in the reference image. As it can be seen in Fig. 1 our proposed method provides superior result compared to the standard method while using same camera poses. The reconstructed building has large number of planar faces at different depths.

The second experiment was performed on 6 images selected from the first half of the BEGIJNHOF sequence with 46 images in total, provided with courtesy of Gallup et al. [6]. The reason we did not employ all the images in the sequence is that we want to demonstrate the method does not require to have a dense sequence. Here and in all other presented experiments we utilize provided camera poses encapsulated in the datasets. This scene is well textured and often used to demonstrate performance of dense stereo algorithms. As can be seen in Fig. 6 our result shows less artifacts than presented in [6] where all images were used.

In the third experiment, we took 6 out of 11 images from the OXFORD CORRIDOR sequence [12]. This scene is extremely difficult for standard dense stereo techniques because the scene contains mostly no-textured surfaces and the images are in gray-scale and in low quality. Despite these problems our method performs very well as shown in Fig. 6.

In the fourth experiment, we use all 5 images from the WADHAM COLLEGE se-

[2] `http://ai.stanford.edu/∼micusik/research.html`

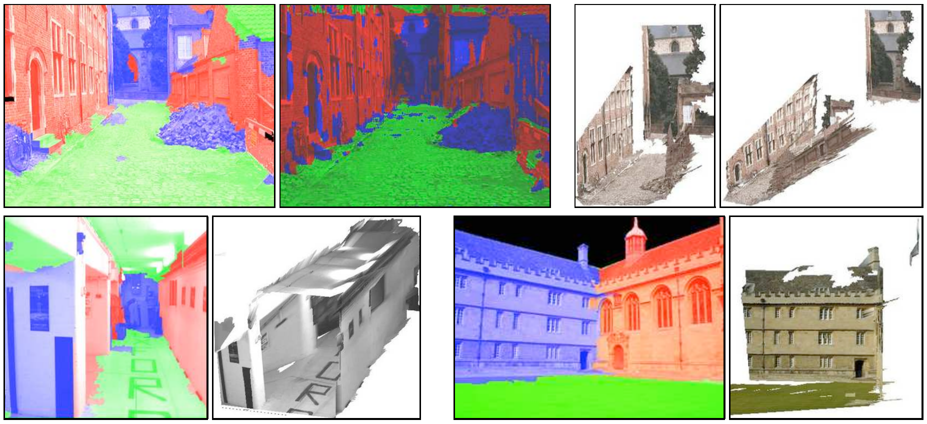

Figure 6: Results. *First row:* The BEGIJNHOF dataset. First two images show normals assigned to pixels by our method and the method by [6]. Notice smoother result by our method. Next two images depict views on 3D reconstruction. *Second row:* The assigned normals and views on 3D reconstructions from the OXFORD CORRIDOR and WADHAMM COLLEGE datasets. *Best viewed in color.*

quence [12]. We show here how the method behaves on surfaces whose normals are not considered in the sweeping stage. Roofs are those examples and as it can be seen in Fig. 6 they are approximated by closest walls while the rest is correctly reconstructed. The frontal view of one of the buildings in the reconstructed 3D model demonstrates correctly estimated planes and shows also a correctly estimated chimney in the corner of the image.

In all our experiments we use the superpixel method based on the Minimum Spanning Tree [4], taking less than 1sec per image, as this method provides more suitable output for our purposes compared to [14]. The energy weights are set to $\lambda_1 = 1$, $\lambda_2 = 0.05$, $\lambda_3 = 0.1$. The performance can be tuned with those parameters, however, even the default setting provides reasonable good performance. The depth sweeping range, i.e. numbers d_{min}, d_{max}, Δd, were set manually for each sequence. This could be possibly avoided by utilizing reconstructed 3D points used for camera motion estimation.

We do not perform a comparison with the Middlebury evaluation database [16] since our method tackles conceptually different scenarios then those present in the benchmark dataset. The focus of the methods is on exploiting constraints of man-made environments and overcoming difficulties of wide-baseline settings.

5 Conclusions

We have presented a method for multi-view 3D dense reconstruction from wide-baseline images of man-made environments. This is a problematic scenario for standard dense stereo methods which often suffers from many artifacts in the final 3D reconstructions. We have shown a natural way to cope favorable with these scenarios by choosing superpixel representation and designing new photometric and geometric cost terms integrated in the MRF framework.

References

[1] H. Cornelius, R. Šára, D. Martinec, T. Pajdla, O. Chum, and J. Matas. Towards complete free-form reconstruction of complex 3D scenes from an unordered set of uncalibrated images. In *Proc. of SMVP Workshop, ECCV*, pages 1–12, 2004.

[2] J.M. Coughlan and A.L. Yuille. Manhattan world: Orientation and outlier detection by bayesian inference. *Neural Computation*, 15(5):1063–1088, 2003.

[3] EosSystems. PhotoModeler, -. `http://www.photomodeler.com`.

[4] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[5] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *Proc. of CVPR*, 2007.

[6] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Proc. of CVPR*, 2007.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[8] K. Kanatani and Y. Sugaya. Statistical optimization for 3-D reconstruction from a single view. *IEICE Trans. on Information and Systems*, E88-D(10):2260–2268, 2005.

[9] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence (PAMI)*, 28(10):1568–1583, 2006.

[10] J. Košecká and W. Zhang. Video compass. In *Proc. of ECCV*, pages 476–490, 2002.

[11] Š. Obdržálek and J. Matas. Object recognition using local affine frames on maximally stable extremal regions. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, LNCS, pages 83–104. Springer, 2006.

[12] Oxford VGG dataset. `http://www.robots.ox.ac.uk/∼vgg/data/data-mview.html`,-.

[13] RealViz. ImageModeler, -. `http://imagemodeler.realviz.com`.

[14] X. Ren and J. Malik. Learning a classification model for segmentation. In *Proc. of ICCV*, pages 10–17, 2003.

[15] A. Saxena, M. Sun, and A.Y. Ng. 3-D reconstruction from sparse views using monocular vision. In *Proc. of VRML Workshop, ICCV*, 2007.

[16] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of CVPR*, pages 519–528, 2006.

[17] Maarten V. and L. Van Gool. Web-based 3D reconstruction service. *Machine Vision Application*, 17(6):411–426, 2006. `http://www.arc3d.be`.

[18] G. Vogiatzis, C. H. Esteban, P. H. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 29(12):2241–2246, 2007.

[19] T. Werner. A linear programming approach to Max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 29(7):1165–1179, 2007.

[20] T. Werner and A. Zisserman. New techniques for automated reconstruction from photographs. In *Proc. of ECCV*, pages 541 – 555, 2002.