Fast Global Labelling For Depth-Map Improvement Via Architectural Priors

Paul Amayo, Pedro Piniés, Lina M. Paz, Paul Newman

***Abstract*— Depth map estimation techniques from cameras often struggle to accurately estimate the depth of large textureless regions. In this work we present a vision-only method that accurately extracts planar priors from a viewed scene without making any assumptions of the underlying scene layout. Through a fast global labelling, these planar priors can be associated to the individual pixels leading to more complete depth-maps specifically over large, plain and planar regions that tend to dominate the urban environment. When these depth-maps are deployed to the creation of a vision only dense reconstruction over large scales, we demonstrate reconstructions that yield significantly better results in terms of coverage while still maintaining high accuracy.**

I. Introduction

With the recent strides in autonomous driving and deployment of robots in large-scale urban environments the ability of robots to create better maps of these environments over large scales have become increasingly important. Dense depth-maps created from monocular/stereo cameras offer a low-cost solution that can naturally deal with both the scale and lighting conditions of outdoor environments but do not have the accuracy of RGB-D cameras which are in turn restricted to small-scale and low-light indoor scenes.

Current state-of-the-art methods for creating dense depth-maps with cameras are based on powerful variational optimisation algorithms [1], [2]. These in general have two terms that are minimised. Firstly a data term that measures the photoconsistency (over a set of consecutive images in the case of a monocular camera or a stereo pair of images) of the depth estimation. Followed by a regularisation term that enforces depth smoothness for homogeneous surfaces while simultaneously attempting to preserve sharp discontinuities between different objects in the scene. A key step of the minimisation process involves the application of a primal-dual optimisation scheme which is widely used for solving variational convex energy functions arising in many image processing problems [3].

The natural challenge in these techniques is dealing with large, plain and planar structures where the data term is of little use, with the lack of texture in these areas restricting the use of photo-consistency. The regularisers promoting smoothness also struggle to propagate information from distant barriers. With urban scenes characterised by a multitude of man-made objects such as roads and buildings which by construction contain planar surfaces this results in noisy and sometimes erroneous depth-maps.

The authors are with the Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK. {pamayo, ppinies, linapaz, pnewman}@robots.ox.ac.uk

Fig. 1. A qualitative perspective of this paper. A dense depth-map (top) of a planar wall surface created by a state-of-the-art Total Generalised Variation algorithm is displayed. As expected the algorithm results in a noisy output on the largely textureless wall and road. Our method discovers planar regions and from there invokes a planar prior to restricted areas. This results in an improved depth-map (bottom), showing a marked improvement in the depth estimation along the wall and road.

In this work we present a vision-only pipeline that accurately extracts and incorporates planar priors into the depth-map estimation through a per-pixel labelling, resulting in more accurate depth-map estimation as shown in Figure 1. With the planar prior extraction and pixel labelling driven by a fast, parallel, global energy minimisation algorithm, this brings real-time capabilities into the depth-map improvement. Planar priors are extracted through a vision-only two view homography segmentation that makes no assumptions about the layout of the scene. To evaluate the performance of the proposed pipeline outlined in Figure 2, the resulting depth-maps were fused to create large-scale dense reconstructions of sections of the KITTI [4] data-set. With this dataset providing ground-truth ego-motion and 3D laser, a quantitative as well as qualitative evaluation of this

proposed depth-map improvement pipeline was undertaken.

In particular this work offers the following contributions:

- A robust and accurate method for extracting planar priors using two-view segmentation.
- A fast global labelling of image pixels to their underlying planes.
- A method to integrate these planar surfaces into large scale dense reconstructions.

With the paper organised as follows. In Section II related work from the literature in improving on depth-map acquisition followed by work on plane prior extraction is presented. In Section III the details of the proposed depth-map estimation technique are described. In Section IV we present the data fusion algorithm used to create the dense reconstructions before showing qualitative and quantitative results in Section V. Conclusions and discussion follow in Section VI.

II. Related Work

The use of strong planar priors in dense depth-map estimation has been explored in several different ways. Most approaches attempt to leverage the powerful variational optimisation machinery by including sophisticated regularisation terms that enforce planarity over the structure. For instance, the work of [5] introduces a non-local higher order regularisation term in a variational framework. This offers significant improvement for large planar surfaces by allowing the propagation of depth information over distant pixels to texture-less regions of the same planar surface.

Other approaches such as [6] use a higher-order term that models Manhattan and piece-wise planar structures. However unlike the cost function proposed by [5], this new regularisation term requires prior estimation of plane normals by using super-pixel classification of indoor Manhattan scenes into predefined classes (wall, floor, ceiling and clutter).

Geometry-only depth-maps are also popular techniques in computer vision. These approaches are motivated by the simplification of scene reconstruction by constraining the reconstruction to geometric surfaces often known as Piecewise Planar Reconstructions (PPRs) [7], [8], [9], [10], [11], [12], [13], as these approaches easily overcome the challenges of poorly textured regions through the strong planar assumption. The depth-map estimation is framed as an optimal labelling problem where each of the pixels is assigned to a particular hypothesised geometric model through a Markov Random Field (MRF) formulation.

Many other approaches generate geometric hypotheses such as those using Manhattan plane models [12], lines and their vanishing points [11], virtual-cut planes [10], plane sweeps [14] and RANSAC [15] all demonstrating good results in their respective application contexts. Most of these approaches, however, neglect non-planar surfaces. [9] mitigates this by adding an extra non planar model represented by the original noisy depth map for those points that are not assigned to any of the hypothesised geometric models. The refinement of the depthmap is formulated as a labelling problem that handles both planar regions through underlying geometric models and non-planar regions. In this

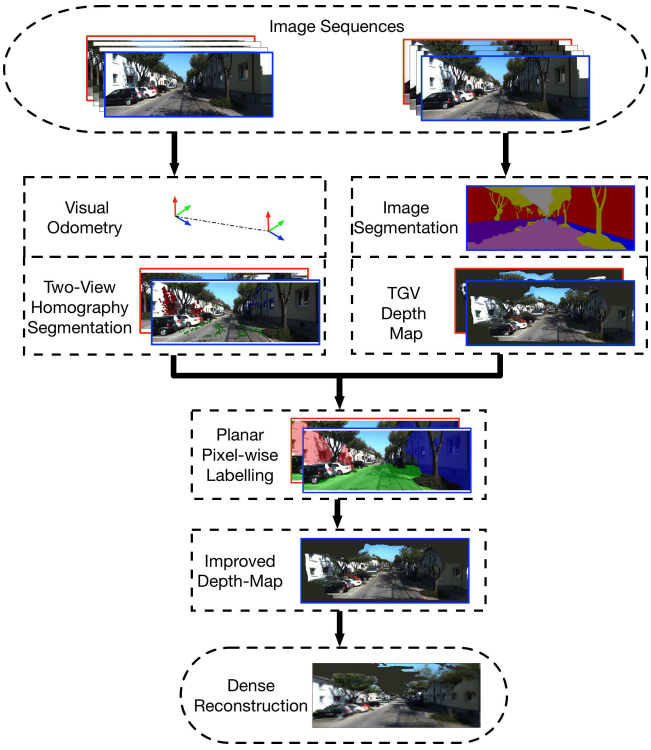

Fig. 2. Overview of the proposed vision only depth-map improvement pipeline. From image sequences (mono or stereo) a two-view homography segmentation in conjunction with visual odometry is used to create planar priors. These priors drive a fast global labelling that assigns planar regions to their corresponding prior and non-planar regions to a depth-map obtained through the Total Generalised Variational (TGV) energy minimization. From this labelling a more complete improved depth-map is estimated. The fusion of consecutive improved depth maps are further fused to create a dense reconstruction.

method, geometric hypotheses are obtained using RANSAC over local connected regions on the original stereo depth-map. While exhibiting good results, this approach is limited to regions of high texture where the depth estimation is fairly accurate which is not the case in plain texture less surfaces. In addition, this method includes a graph-cut labelling based on the α-expansion [16], a sequential algorithm that cannot exploit parallelisation in modern General Purpose GPUs [17] whose use is desirable to ensure real-time execution.

The work of [18] describes a continuous optimisation approach allowing for real-time optimal labelling. By utilising a first-order energy minimisation algorithm, this approach lends itself to a highly parallel GPU implementation. A speed-up of around 30 times is reported as compared to the discrete graph-cut approach. However to initialise the geometric models a plane sweep algorithm is developed [14] restricting the use to layouts where the ground plane and two orthogonal facades are dominant. This is an approximation to a Manhattan world, an assumption that does not hold throughout the urban scene leading to errors when new planar configurations are viewed. Moreover, this method does not effectively deal with non-planar regions.

In this work we opt for an initialisation of geometric priors that does not make any assumptions about the underlying lay-

out of the scene. We extend the work of [5] to estimate depth-maps with non-local Total Generalised Variation (TGV) term to favour planar surfaces in textureless regions. Unlike [18] we do not limit planar priors to only ground plane configurations. We propose a solution for automatic planar prior discovery by means of a two-view multi-homography segmentation of the scene. To accurately detect multiple homographies in the presence of high levels of noise and clutter is, however, a non-trivial task for classical approaches such as RANSAC [15]. In contrast, robust multi-model fitting approaches such as [19] have shown to significantly outperform greedy clustering based approaches under high levels of noise by means of a global discrete energy minimisation. In this work, we use a similar formulation in terms of a global continuous energy minimisation with primal dual algorithm. The resulting homographies can be decomposed to reveal the underlying planar priors. As in [9], we use an extra label to account for non-planar models, together with colour image segmentation. Given this information we refine the original stereo depthmaps in a fast per-pixel depth labelling with parallel continuous energy minimisation. We show that our approach can be used to produce volumetric reconstructions of urban environments with significant improvements in terms of completeness of the surface without compromising accuracy.

III. Depth-Map Estimation with Planar Priors

In this Section we describe the estimation of the depthmaps in a labelling framework that leverages of automatic planar prior discovery. While this pipeline can be applied to monocular or stereo cameras we focus here on a stereo implementation.

A. Stereo Depth-Map Initialization

The first component of the pipeline is a module that produces stereo disparity maps from which our depthmaps are obtained. The disparity estimation algorithm solves the following variational problem:

$$\min_{\zeta} E_{reg}(\zeta) + E_{data}(\zeta). \tag{1}$$

The data term measures the photoconsistency between corresponding pixels in the stereo images. This is given by

$$E_{data}(d; I_L, I_R) = \iint_{\Omega} |\rho(d, x, y)| \mathrm{d}x\mathrm{d}y. \tag{2}$$

where (x,y) are the coordinates of a pixel in the reference image, and the function $\rho(d, x, y) = Sim^W(IL(x + d, y), IR(x, y))$ measures the similarity between two pixels using a patch window W for a candidate disparity $d \in D$.

We then reach for a Total Generalised Variation (TGV) regularisation term which favours planar surfaces in any orientation:

$$E_{reg}(d) = \min_{w \in \mathbb{R}^2} \alpha_1 \iint_{\Omega} |\mathbf{T}\nabla d - \mathbf{w}| \mathrm{d}x\mathrm{d}y + \alpha_2 \iint |\nabla \mathbf{w}| \mathrm{d}x\mathrm{d}y. \tag{3}$$

where $\mathbf{w}$ allows the disparity d in a region of the depth map to change at a constant rate and therefore creates planar surfaces with different orientations and $\mathbf{T}$ is a tensor that preserves object discontinuities. This tensor is included to mitigate the tension between maintaining object discontinuities and keeping smoothness in the energy minimisation. While the appearance gradient (∇I) can indicate the presence of boundaries between objects, it does not include any information on the direction of these borders. To take this information into account we adopt an anisotropic diffusion tensor:

$$\mathbf{T} = exp(-\gamma |\nabla \mathbf{I}|^{\beta}) nn^T + n^{\perp} n^{\perp^T} \tag{4}$$

where $n = \frac{\nabla I}{|\nabla I|}$ and $n^{\perp}$ is its orthogonal complement. $\mathbf{T}$ decomposes the disparity gradient (∇d) in directions aligned with n and $n^{\perp}$. We penalise components aligned with $n^{\perp}$, but do not penalise large gradient components aligned with n, such as those appearing due to lighting changes. In other words, if there is a discontinuity visible in the colour image, then it is highly probable that there is a discontinuity in the depth image.

The minimisation of this energy is then performed by an iterative exhaustive search step for the non-convex data term E_{data} and a Primal-Dual algorithm [3] for the convex regularisation term [20] E_{reg}. From the resulting disparity-map the well defined camera intrinsic matrix $\mathbf{K}$ can be used to create the stereo depth-map.

B. Planar Prior Generation

Although planar structures in urban scenes are mainly represented by texture-less regions, we can leverage of the existence of objects that can be revealed by blob detectors such as SURF. We leverage of this insight to extract sparse features from the scene. Then we follow the classical multiple-view framework to find correspondences between sparse points across two views. We use the relation between point correspondences and an observed 3D plane through the well established homography. Without lost of generality, given a sparse set of n pixel correspondences in homogeneous coordinates between the two views $\mathbf{u}_i = (\mathbf{u}_i^1, \mathbf{u}_i^2) \in \mathbb{R}^2,\ i = 1 \cdots n$, a homography $\mathbf{H}^{21} \in \mathbb{R}^{3\times3}$ establishes the mapping of pixels from the first view $\mathbf{u^1}$ to the second view $\mathbf{u^2}$ through an observed plane π with normal vector $\mathbf{n}^1$ and distance d^1 [21]. With this available information, we can extract the motion between the two views $(\mathbf{R}^{21}, \mathbf{t}^{21})$.

$$\mathbf{u}^2 = \mathbf{H}^{21}\mathbf{u}^1, \tag{5}$$

$$\mathbf{H}^{21} = \mathbf{R}^{21} + \frac{\mathbf{t}^{21}(\mathbf{n}^1)^T}{d^1}. \tag{6}$$

Similar to [22] and [23] we initialise our homography fitting problem from affine transformations to achieve better performance as compared to the classical Direct Linear Transform [24]. For this initialisation a non homogenous point correspondence $\mathbf{u}_i$ is augmented by the 2 X 2 affine matrix $\mathbf{A}$ that maps the image points surrounding $\mathbf{u}_i^1$ into those in the vicinity of $\mathbf{u}_i^2$ [22].

$$\mathbf{A} = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix} \tag{7}$$

Two affine correspondences $(\mathbf{u}_i, \mathbf{A})$ and $(\mathbf{u}_j, \mathbf{B})$ then belong to the same homography if they satisfy the following

conditions.

$$
\begin{aligned}
(\mathbf{u}_j^2 - \mathbf{u}_i^2)^T \mathbf{P}\mathbf{A}(\mathbf{u}_j^1 - \mathbf{u}_i^1) = 0, \\
(\mathbf{u}_j^2 - \mathbf{u}_i^2)^T \mathbf{P}\mathbf{B}(\mathbf{u}_j^1 - \mathbf{u}_i^1) = 0, \\
\begin{bmatrix} s + a_2b_3 - a_3b_2 & -(a_1b_3 - a_3b_1) \\ a_2b_4 - a_4b_2 & s - (a_1b_4 - a_4b_1) \end{bmatrix} (\mathbf{u}_j^2 - \mathbf{u}_i^2) = \mathbf{0}.
\end{aligned} \tag{8}
$$

Where the variables s and $\mathbf{P}$ are defined as

$$
s = \frac{[-a_2 + b_1a_1 - b_1](\mathbf{u}_j^2 - \mathbf{u}_i^2) - (a_1b_2 - a_2b_1)(x_i^1 - x_j^1)}{(y_i^2 - y_j^2)}
$$

$$
\mathbf{P} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.
$$

The average error of the four conditions provides a similarity score between pairs of point correspondences. We cluster point correspondences belonging to the same homography using affinity propagation [25]. Noise and outlier removal is performed in a refinement step for which the quality of the solution is linked to the global energy minimisation. In this process pixel correspondences are used to fit homography models while simultaneously considering the overall classification of points.

This gives the global energy as in Equation 9:

$$
\underbrace{\sum_{l=1}^{L} \left(\sum_{i=1}^{n} (\|D(\mathbf{u}_i^1, \mathbf{H}_l^{12}\mathbf{u}_i^2)\|_{\Sigma_{12}} + \|D(\mathbf{u}_i^2, \mathbf{H}_l^{21}\mathbf{u}_i^1)\|_{\Sigma_{21}} \phi_l(\mathbf{u}) \right)}_{\text{Data Term}}
$$

$$
+ \lambda \underbrace{\sum_{l=1}^{L} \left(\sum_{i=1}^{n} \omega_{\mathcal{N}} |\nabla_{\mathcal{N}} \phi_l(\mathbf{u})|_{1,1} \right)}_{\text{Smoothness Term}} + \underbrace{\beta L}_{\text{Label Term}} . \tag{9}
$$

The data term in Equation 9 accounts for the symmetric transfer of the re-projection error. Here, we refer to D as the Mahalanobis distance $\|D(\mathbf{u}_i, \mathbf{H}_{ab})\|_{\Sigma_{ab}} = (\mathbf{u}_i^a - \mathbf{u}_i^{a'})\Sigma_{ab}^{-1}(\mathbf{u}_i^a - \mathbf{u}_i^{a'})^T$ where Σ_{ab} represents the propagated covariance matrix through the mapping induced by the corresponding homography.

The assignment of data points to their respective models is encapsulated through an indicator function

$$
\phi_l(\mathbf{u}) = \begin{cases} 1 & \mathbf{u} \in L_l \\ 0 & \text{otherwise} \end{cases} \tag{10}
$$

where the uniqueness in the label assignment can be achieved by adding the constraint $\sum_{l=1}^{L} \phi_l(\mathbf{u}) = 1$. To account for outliers –where some data points might not be explained by a geometric model– a special label $\emptyset$, representing the outlier model is added. In this way a constant cost is assigned γ to points that cannot be explained by any geometric model. The model cost for the outlier model is simply given by $\rho_\emptyset(\mathbf{u}, \phi_\emptyset(\mathbf{u})) = \gamma$.

The smoothness term in Equation 9 takes into account locality by promoting a homogeneous assignment of labels to neighbouring points. The $\nabla_{\mathcal{N}}$ operator calculates the gradient of the indicator function over the neighbourhood $\mathcal{N}$ of a point given in this work by its k-nearest neighbours.

Algorithm 1: Multi-Homography fitting through global energy algorithm.

```
Initialisation;
Propose L models;
while not converged do
    Primal Dual Optimisation;
    Merge Homographies;
    Re-estimate Homographies;
end
```

This penalises points that belong to the same neighbourhood but do not share the same model. The parameter λ controls the trade-off between the smoothness cost and the data cost while the weights $\omega_{\mathcal{N}}$ are used to reduce the effect of the smoothness term depending on the distance between a point and its nearest neighbour.

Finally, the third term in Equation 9 penalises the number of models by adding a constant cost β per model. This eliminates redundancies in models resulting in a compact solution.

The minimisation of this energy can be performed using a discrete optimisation algorithm such as α-expansion [16]. For this work we prefer a continuous optimisation approach outlined in [26] that leverages a primal dual optimisation [3] to perform the energy minimisation. This approach by utilising a parallel approach implementable on a GPGPU is able to achieve faster execution time as compared to the discrete α-expansion approach. Allowing for real-time performance on geometric model detection. The multi-homography fitting algorithm is shown in Algorithm 1 and we refer the reader to [26] for further implementation details.

After algorithm 1 converges, N homographies are retrieved. The underlying plane priors defining the homographies are extracted by applying Singular Value Decomposition (SVD) [27] on $\mathbf{K}^{-1}\mathbf{H}^{\mathbf{12}}\mathbf{K}$ with $\mathbf{K}$ describing the camera intrinsics. Then we extract the motion $\{\mathbf{R}, \mathbf{t}\}$ and the plane parameters $\{\mathbf{n}, d\}$ described by Equation 6.

Notice that SVD leads to two valid separate solutions for $\{\mathbf{R}, \mathbf{t}, \mathbf{n}\}$. In order to disambiguate between the two, we use the ego-motion estimation $\mathbf{T}_{vo}$ from our visual odometry,

$$
\mathbf{T}_{vo} = \begin{bmatrix} \mathbf{R}_{vo} & \mathbf{t}_{vo} \\ \mathbf{0} & 1 \end{bmatrix} \tag{11}
$$

Additionally, this decomposition only gives the translation and distance of the plane upto scale. The ego-motion translation estimate from visual odometry is therefore used to obtain the actual distance of the plane as follows,

$$
d_a = \frac{\mathbf{t}_{vo}}{\mathbf{t}} \tag{12}
$$

C. Semantic Image Segmentation

Before we carry out any depthmap refinement, an extra pre-processing step is added to improve the depth accuracy. In other words, certain classes of objects found in the urban scene are known to be non-planar including cars, trees and pedestrians that only account as noise to the per-pixel depth labelling. Therefore similar to the approach presented in [9],

a semantic image segmentation is included to determine the class of pixels in the scene. This will help to exclude non-planar object classes from the planar labelling.

For this work, a full resolution residual network as outlined in [28] is used. This approach combines outstanding recognition performance, as found in the current state-of-the-art with increased localisation accuracy. Training data was obtained from manually annotated images in the KITTI dataset [29] with an example output shown in Figure 3.

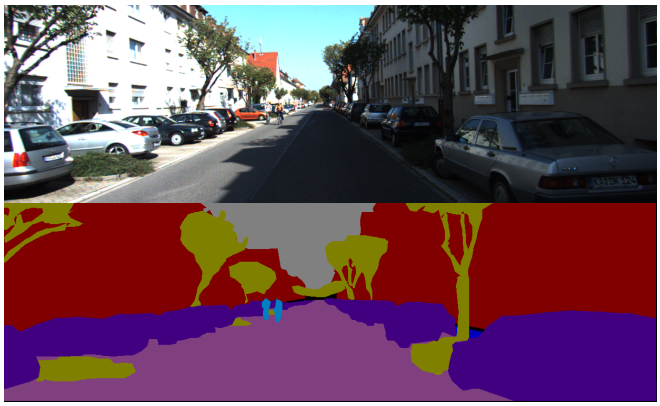

Fig. 3. Semantic Segmentation of the pixels of an image (top) into its respective classes (bottom). Buildings and roads over which planar surfaces are expected to be found are labelled in red and pink respectively.

D. Fast Global Labelling

Once the planar priors are retrieved from the two-view multi-homography segmentation, we propose a labelling of the image pixels to the corresponding underlying planes. Analogous to the homography minimisation, we use a similar global energy approach that leverages of the detected planar priors. Our energy is formulated as

$$\sum_{l=1}^{L}\Bigg(\underbrace{\sum_{i=1}^{n}\|D(\mathbf{u})\|\phi_l(\mathbf{u})}_{\text{Data Term}}+\lambda\underbrace{\sum_{i=1}^{n}\omega_{\mathcal{N}}\|\nabla_{\mathcal{N}}\phi_l(\mathbf{u})\mathbf{S}\|_p}_{\text{Smoothness Term}}\Bigg) \quad (13)$$

Where L is the number of plane hypotheses with an addition of the non-planar label $\emptyset$ that is provided by the original TGV stereo depth-map, with the nodes $\mathbf{u} \in D$ the image pixels. With $\phi(\mathbf{u})$ being the indicator function the data term is defined in Equation 14.

$$D(\mathbf{u}_l) = \begin{cases} \min(\rho(\mathbf{u}_l), \rho_{min}) & l \in \pi_i, \cdots, \pi_L \\ \min(\rho(\mathbf{u}_\emptyset), \rho_{min}) + \rho_{bias} & l_\emptyset \end{cases} \quad (14)$$

$\rho(\mathbf{u})$ is the photo-consistency measured between images. At each frame in this work there are four images available, the current stereo pair coupled with a previously viewed stereo pair from which the two view homography segmentation is calculated. With the baseline and ego-motion known, pixels can be projected from an image to either of the remaining three according to their corresponding depths. This depth is determined by the planar priors or given by the TGV stereo depth map in the outlier case as shown in Equation 14. From this projection, the photo-consitency is measured by comparing the similarity of the projected pixels and that of the second image over a window of size W. Poorly matched surfaces are handled through ρ_{min}, while a small bias term ρ_{bias} is added to the outlier label.

Additionally, to ensure that non-planar regions are not wrongly assigned into planar priors, pixels that belong to non-planar classes have $\rho(\mathbf{u})$ automatically amended w.r.t. the semantic segmentation as shown in Equation 15. This semantic segmentation include cars, trees and pedestrian classes.

$$\rho(\mathbf{u}_l) = \begin{cases} \rho_{max} & Non-Planar \\ \rho(\mathbf{u}_l) & Planar \end{cases} \quad (15)$$

The smoothness term penalises depth discontinuities between neighbouring pixels based on their plane labels. A 4-connectivity pattern is used for $\mathcal{N}$ with a Euclidean norm to penalise points that belong to the same neighbourhood but have different labels. $\mathbf{S}$ is a matrix of the pixel depths corresponding to their assignment to different labels. To preserve depth discontinuities arising from different objects the smoothness term was down-weighted based on changes in the image gradient (∇I) using the weighting function $\omega_{\mathcal{N}}$ presented in Equation 16.

$$\omega_{\mathcal{N}} = \frac{1}{\gamma\nabla I^2 + 1} \quad (16)$$

As in the case of the planar prior generation, a continuous optimisation approach is opted for [26]. This not only has time benefits but also allows for different norms to be used. The metrication error arising from graph-cut techniques [17] can thus be avoided as an Euclidean norm is used. From results presented in Figure 4 it can be seen that this approach is able to perform accurate labelling over different scenes.

E. Implementation

We implemented and tested this pipeline using CUDA on a Nvidia GeForce GTX TitanBlack 6048MB GPU. Running it 100 times over sample images from KITTI [4] with sample results shown in Table I. The images are of resolution 376x1241. Bench-marking was run in scenes with 3 planar models mirroring the most common application scenario.

TABLE I
TIMING RESULTS ON THE DEPTH-MAP IMPROVEMENT PIPELINE

Module	Time (ms)
Homography Optimisation (≈1000 correspondences)	13.7
Photo-consistency computation	3.7
Per-pixel labelling	185.3

An α-expansion approach run on the per pixel labelling problem returns a running time of around three and a half seconds for the images presented, as can be seen in Table I the proposed pipeline is able to report a much faster labelling (around twenty times). This reported time is achieved by using a coarse to fine approach that reduces the number of iterations needed to converge as demonstrated in [18]. This is done by performing the labelling at various levels: the original resolution, half the resolution and a quarter the resolution. With this approach a reduction of the number of iterations is achieved up to a factor of ten.

Fig. 4. By using a global energy approach, our proposed approach is able to detect multiple homographies in images as shown in the top row. These homographies encode planar priors which are used as input into the planar labelling of the image pixels. The bottom row shows the results of the pixel labelling using the extracted planar priors.

IV. Dense Reconstruction

A. Depth Map Fusion

To create a dense reconstruction in this work the BOR^2G approach outlined by [30] is followed. In this work, the incoming depth values are processed through a uniform voxel grid. Each voxel stores range observations represented by their corresponding Truncated Signed Distance Function (TSDF), u_{TSDF}. The voxels TSDF values are computed such that one can solve for the zero-crossing level set (isosurface) to find a continuous surface model. Even though the voxel grid is a discrete field, because the TSDF value is a real number, the surface reconstruction is even more precise than the voxel size.

To deal with the large spatial region represented in these reconstructions, a hashing voxel grid (HVG) is used [31]. This subdivides the world into an infinite and uniform grid of voxel blocks, which in turn encapsulate their own small voxel grid. Only when a surface in a given voxel block is observed, are all the voxels in that grid allocated and their corresponding TSDF values updated. Applying a hash function to coordinates in world space gives an index within the hash table, which in turn points to the raw voxel data.

For each of the voxel blocks, the update equations are identical to those presented by [32] which projects voxels into the depth map to update the TSDF data (f). The fusion step can then be posed as a noise-reduction problem that can be approached by a continuous energy minimisation over the voxel-grid domain (Ω):

$$E(u) = \int_{\Omega} |\nabla u| \mathrm{d}\Omega + \lambda \int_{\Omega} ||f - u||_2^2 \mathrm{d}\Omega \qquad (17)$$

where E(u) is the energy (which we seek to minimise) of the denoised (u) and noisy (f) TSDF data. The regularisation energy term, commonly referred to as a Total Variation (TV) regulariser, seeks to fit the solution (u) to a specified prior the L1 norm in this case. The data energy term seeks to minimise the difference between the u and f , while λ controls the relative importance of the data term vs. the regulariser term. For further implementation details we direct the reader to [33].

V. Results

In this section the proposed pipeline over the dense reconstruction task is evaluated using the publicly available KITTI dataset [4]. Comparing the reconstructions produced using the raw depth-maps and those improved through the planar priors.

To create the ground truth for the comparison, the laser scans from the Velodyne HDL-64E were consolidated into a single reference frame in a similar approach to Tanner et al. [33]. As there are inevitable errors in the KITTI's GPS/INS ground -truth poses we limited this consolidation to a relatively short distance (75m) over which a good comparison between the two could still be made. To compute the error statistics against the sparse ground cloud point cloud, the vertices of the corresponding mesh were used. The error thus reduced to the metric distance between a vertex in the mesh and its closest laser point. Two sections of the KITTI sequence 00 were chosen for evaluation in regions where the reconstruction performance was shown to be poor. Using ten cm voxels the reconstructions were created from the two sets of depth maps with results consolidated in Table II and Figure 5.

From Figure 5 it can be seen that the introduction of the planar priors results in more complete reconstructions (bottom row) over the evaluated sections than the raw TGV depth-maps (top row). Holes in the reconstruction created as a result of the noisy depth maps in textureless regions are filled as well as gaps presented through partially occluded region through the planar prior. Resulting in a smoother and more complete reconstruction.

This is confirmed when the quantitative evaluation in Table II is viewed. For both cases the use of the planar prior results in a significant increase in the surface area of the reconstruction, $\approx$ 10 % in the first case, as well as number of blocks and voxels when the planar priors are used. This can be expected as the resulting reconstruction covers a larger area and thus accumulates more error. We do additionally observe in both cases that the difference between the median errors is within the noise of the ground truth laser sensor (3 cm).

From this it can be observed that the proposed pipeline does successfully improve the depth map estimation of large, planar textureless surfaces over our test cases. Thus allowing low cost visual sensors to more completely capture the urban environment while still remaining within the error margins.

Fig. 5. Sample results of dense reconstructions from the fusion of depth maps in the two evaluated sections are shown in this figure. The images on the left correspond to the reconstructions produced using the TGV depth maps. It can be seen that the noise on the depth maps results in some gaps in the reconstruction. By improving these depth maps with planar priors in our approach the results can be improved on significantly as shown in the images on the right. Resulting in a more accurate map of the world.

TABLE II

TABLE OF THE EVALUATED ERROR STATISTICS.

KITTI-VO 00	Type	Median (cm)	75% (cm)	Surface Area (m^2)	#Blocks	#Voxels 10^6	GPU Memory (MB)	Distance (m)
Case 1	TGV	15.9434	49.8064	3140.74	26367	13.49	154.494	78.22
—	Prior	18.2147	51.5005	3450.91	27405	14.03	160.576	78.22
Case 2	TGV	24.4837	99.8326	4162.02	28185	14.43	165.146	73.67
—	Prior	25.6882	98.3273	4399.76	30274	15.50	177.387	73.67

VI. CONCLUSIONS

In this work an end to end system for improving the estimation of depth-maps from images is presented and evaluated. In particular focusing on large, plain and planar surfaces which have long plagued contemporary dense depth-map estimation techniques due to lack of abundant texture. We leverage an energy based model discovery technique over homographies to induce planar priors over these planar regions. This is then similarly employed to associate and label the underlying image pixels to their corresponding planar model to improve the depth-map estimation. The labelling is further aided by an image segmentation that removes the pixels that belong to non-planar regions, reducing the noise and increasing the labelling accuracy.

The energy minimisations in this work rely on a continuous optimisation approach CORAL [26], that allows for a parallel implementation on GPGPU hardware. This when run on images from the KITTI dataset [4] resulted in a labelling that was twenty times faster than other discrete optimisation approaches, most notably α-expansion [16]. Opening up this technique to online use in robotics applications and with steady advances in GPU hardware real-time implementations. This would reduce the reliance on expensive 3D sensors in favour of lower-cost visual sensors allowing for the increased development and deployment of autonomous vehicles. The subsequent fusion of the improved depth-maps also reveals a more complete reconstruction of the environment with little change in the reconstruction error.

VII. Acknowledgements

The authors acknowledge the following funding sources. Paul Amayo is funded by the Rhodes Trust. Paul Newman Lina Paz and Pedro Pinies is are/were supported by EPSRC Programme Grant EP/M019918/1.

References

[1] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Joint Pattern Recognition Symposium*. Springer, 2010, pp. 11–20.

[2] G. Graber, T. Pock, and H. Bischof, "Online 3d reconstruction using convex optimization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 708–711.

[3] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] P. Pinies, L. M. Paz, and P. Newman, "Dense mono reconstruction: Living with the pain of the plain plane," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5226–5231.

[6] A. Concha, M. W. Hussain, L. Montano, and J. Civera, "Manhattan and piecewise-planar constraints for dense monocular mapping." in *Robotics: Science and systems*, 2014.

[7] C. Raposo, M. Antunes, and J. P. Barreto, "Piecewise-planar stereoscan: structure and motion from plane primitives," in *European Conference on Computer Vision*. Springer, 2014, pp. 48–63.

[8] J. Molnar, R. Huang, and Z. Kato, "3d reconstruction of planar surface patches: A direct solution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 286–300.

[9] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1418–1425.

[10] M. Antunes, J. P. Barreto, and U. Nunes, "Piecewise-planar reconstruction using two views," *Image and Vision Computing*, vol. 46, pp. 47–63, 2016.

[11] S. N. Sinha, D. Steedly, and R. Szeliski, "Piecewise planar stereo for image-based rendering," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1881–1888.

[12] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1422–1429.

[13] ——, "Reconstructing building interiors from images," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 80–87.

[14] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[17] C. Nieuwenhuis, E. Toeppe, and D. Cremers, "A survey and comparison of discrete and continuous multi-label optimization approaches for the potts model," *Int. Journal of Computer Vision*, vol. 104, no. 3, pp. 223–240, 2013.

[18] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer, "Fast global labeling for real-time stereo using multiple plane sweeps." in *VMV*, 2008, pp. 243–252.

[19] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *International journal of computer vision*, vol. 97, no. 2, pp. 123–147, 2012.

[20] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.

[21] A. Agarwal, C. Jawahar, and P. Narayanan, "A survey of planar homography estimation techniques," *Centre for Visual Information Technology, Tech. Rep. IIIT/TR/2005/12*, 2005.

[22] C. Raposo and J. P. Barreto, "Theory and practice of structure-from-motion using affine correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5470–5478.

[23] D. Barath and L. Hajder, "Novel ways to estimate homography from local affine transformations," in *Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2016, pp. 432–443.

[24] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[25] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[26] P. Amayo, P. Pinies, L. M. Paz, and P. Newman, "Geometric multi-model fitting with a convex relaxation algorithm," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

[27] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 03, pp. 485–508, 1988.

[28] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," *arXiv preprint arXiv:1611.08323*, 2016.

[29] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. M. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 231–238.

[30] M. Tanner, P. Piniés, L. M. Paz, and P. Newman, "BOR2G: Building Optimal Regularised Reconstructions with GPUs (in cubes)," in *International Conference on Field and Service Robotics (FSR)*, Toronto, ON, Canada, June 2015.

[31] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 169, 2013.

[32] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[33] M. Tanner, P. Piniés, L. M. Paz, and P. Newman, "Keep geometry in context: Using contextual priors for very-large-scale 3d dense reconstructions," in *Robotics: Science and Systems, Workshop on Geometry and Beyond: Representations, Physics, and Scene Understanding for Robotics*, June 2016.