Exploiting high level scene cues in stereo reconstruction

Simon Hadfield
University of Surrey
Guildford, UK, GU2 7XH
s.hadfield@surrey.ac.uk

Richard Bowden
University of Surrey
Guildford, UK, GU2 7XH
r.bowden@surrey.ac.uk

Abstract

We present a novel approach to 3D reconstruction which is inspired by the human visual system. This system unifies standard appearance matching and triangulation techniques with higher level reasoning and scene understanding, in order to resolve ambiguities between different interpretations of the scene. The types of reasoning integrated in the approach includes recognising common configurations of surface normals and semantic edges (e.g. convex, concave and occlusion boundaries). We also recognise the coplanar, collinear and symmetric structures which are especially common in man made environments.

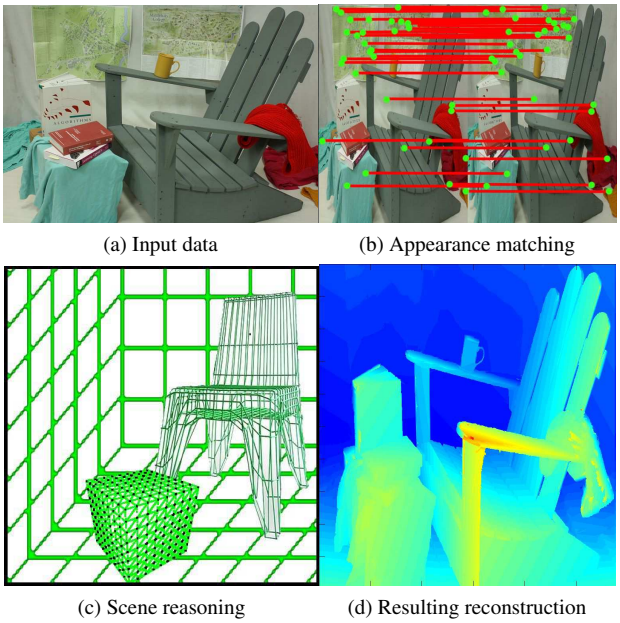
1. Introduction

Understanding the 3D structure in an environment purely from visual observations is one of the oldest and most widely exploited problems in computer vision. It is also one of the most challenging problems for general scenes; many ambiguities result from different combinations of structure, texture and illumination leading to the same observed images. We present a novel formulation for the problem, which makes it possible to unify both bottom-up appearance matching and top-down scene reasoning, in a single approach.

This formulation is inspired by the human visual system. Recognising matches between the observations of both eyes allows depth to be estimated via triangulation. This (along with assumptions about the smoothness of the scenes structure) can be seen as the traditional approach to stereo reconstruction dating back as far as the 1960s [2, 22]. In computer vision this is generally achieved by estimating the epipolar geometry (equivalent to a humans innate knowledge of their eyes characteristics) followed by some form of appearance based matching. In this paper we refer to this as bottom-up reconstruction, as the reconstruction emerges from the matching of small scene sub-units.

However, humans also use many strong high-level cues to understand the structure of their environment. This can be seen intuitively, by noticing that people can easily understand the layout of objects within a photo or video, even though triangulation indicates that all objects lie on a single plane. Many of these cues have been explored in computer vision, particularly when working from a single image (e.g. single-image reconstruction and scene understanding). Some of the most commonly used cues are assumptions about the viewing orientation and gravity, assumptions about the type of surfaces in manmade-environments (i.e. the Manhattan world assumption) and assumptions about commonly occurring object configurations. We collectively refer to these as top-down reconstruction techniques, as the structure of each scene sub-unit is defined using rules about the overall configuration of multiple sub-units.

One particular advantage of our unified framework is it reduces issues related to the baseline of standard stereo systems. Matching and triangulation based systems tend to



(a) Input data (b) Appearance matching

(c) Scene reasoning (d) Resulting reconstruction

Figure 1: Illustration of bottom-up matching (b) and top-down understanding (c) cues used in our unified approach.

only be accurate at distances similar to the separation of the cameras. Beyond this range, small errors in the triangulation will manifest as large errors in depth. In contrast, the unified system is able to smoothly transition to top-down reconstruction as bottom-up becomes less reliable and mirrors the behaviour of human depth, where researchers have discovered that different cues have different operating ranges [6]. This results in 3 general "perceptual spaces": the near space (where triangulation is the dominant reconstruction cue), the ambient space (which uses a combination of bottom-up and top-down reconstruction), and the vista space which relies almost exclusively on top-down information. Our approach smoothly interpolates between these 3 states.

2. Related Work

2.1. Bottom-up reconstruction

Bottom-up approaches to reconstruction are based on matching and triangulation between different viewpoints of the same scene. Local approaches to bottom up reconstruction are based on independent matching between sets of distinctive feature points. The most prevalent of these approaches is matching based on feature points such as the SIFT-descriptor [28]. Recently, many more advanced matching criteria have been proposed including: edge preserving filters [29], generative models [10] and the census transform [21].

These local matching approaches can be limited to operate along epipolar lines for calibrated reconstruction, or can operate over the whole scene in order to estimate the calibration [26]. Recently, the semi-global matching approach originally proposed by Hirschmuller [19] has become a particularly popular extension to epipolar search, due to improved accuracy and robustness to calibration inaccuracies. Recent contributions in this area include iterative [18] and weighted [33] semiglobal matching.

Regardless of the matching technique, purely local reconstruction cannot operate on general scenes. Due to the aperture problem, extracted descriptors cannot be reliably matched in regions which do not have strong texture perpendicular to the epipolar line. Because of this, most recent work in bottom-up reconstruction focuses on global approaches, which combine local matching costs with various spatial smoothness constraints. Various approaches to encoding these spatial smoothness have been proposed including: Total Variation (L2 and L1 [39, 24]), Monte-carlo inspired PatchMatch approaches [3, 17] and the Total Generalized Variation [31, 25] which helps overcome the staircasing artifacts caused by total variation regularisation.

Another approach to encouraging local smoothness is to build the reconstruction out of primitives, rather than estimating a depth for each pixel. This is the standard approach for top-down scene understanding, but has also been exploited in bottom up reconstruction. At the simplest level, oriented planes are used as reconstruction primitives [37]. More detailed reconstructions may be achieved by using curved surfaces, at the cost of increased matching difficulty [40]. In earlier work, the most complex level of reconstruction primitives were geometric subunits or "geons" [36], but more recently these have been replaced by whole-object primitives [5, 4].

2.2. Top-down reconstruction

Most often, top-down reconstruction employs these primitive sub-units, and formulates constraints on the relationship between sub-units. At the local level, frequent relationships between small numbers of neighbouring oriented planes [8] and concave/convex edges [9] can provide a great deal of information about the scene. This idea has recently been extended to exploit a learned representation via convolutional neural networks [34]. This idea can be extended to interpreting different types of relationship between groups of primitives, such as "on-top-of", "supporting", "occluding" etc. [12].

Global top-down constraints have tended to focus on exploiting properties of man-made environments, e.g. room interiors may be coarsely modelled as the inside of cuboids with 3-5 visible faces [16, 15]. Perhaps the most common global top-down constraint is the Manhattan-world assumption (that scenes are composed of planes from only 3 orthogonal directions) [27].

2.3. Joint approaches

There has been a small amount of work which attempts to combine bottom-up and top-down reconstruction techniques. These approaches tend to follow the "Reconstruction meets Recognition" paradigm, where an initial detection stage is included, to locate a set of pre-determined classes, which then inform reconstruction. For specialised categories of environment, this can prove extremely effective, for example Hane *et al*. [13] reconstruct urban scenes by differentiating buildings, sky, ground, vegetation and clutter. Each class then has associated weightings favouring different types of reconstruction. Very recently, Guney and Geiger [11] have forgone this weighting procedure, instead using detection of cars in driving footage to transfer 3D car models into the reconstruction.

The drawback of these approaches is they are only effective in a particular class of environment, limited by the classes which the recognition pipeline is trained for. In contrast, we propose a formulation which integrates far more top-down cues, while also avoiding the need for class specific learning. In addition, unlike these 2 stage procedures, the joint formulation inherently balances top-down and bottom-up information based on the configuration of

the cameras and content of the scene.

3. Unified bottom-up and top-down reconstruction

We will next introduce the representation used to enable effective fusion of bottom-up and top-down reconstruction cues (Section 4). We will then introduce a number of bottom-up reconstruction cues within this framework (Sections 4.1 and 4.2. Various types of top-down scene knowledge will then be introduced and integrated into the system in Section 5. Section 6 describes the efficient optimisation scheme developed for this task, and Section 7 evaluates it on the recent Middlebury 2014 [32] benchmark.

4. Bottom up reconstruction

We formulate our reconstruction in terms of primitives, which is a common technique in both the bottom-up and top-down literature. A set of superpixels $\mathcal{S}$ is extracted from the reference image I^r. Note that we do not attempt to extract and match against superpixels from the target image I^t, this is because superpixel segmentation is not robust to viewpoint changes, particularly for wide baseline stereo.

Instead each superpixel ($s_i \in \mathcal{S}$) is parametrised as an oriented plane primitive, and these primitives are used to perform matching directly between the two images. The parametrisation of each plane is a vector ($\boldsymbol{\alpha}_i \in \mathbb{R}^3$) which corresponds to the normal vector of the plane, divided by the perpendicular distance to the plane. Using this representation, any point ($\mathbf{p} \in \mathbb{R}^3$) which lies on the plane satisfies the condition $\boldsymbol{\alpha}^\top \mathbf{p} = 1$. Furthermore, given the direction vector of any ray $\mathbf{r}$, the distance along that ray at which it intersects the plane is given by $d = 1/(\mathbf{r}^\top \boldsymbol{\alpha})$.

With two static cameras, an oriented planar surface induces a homography between the two images. Without loss of generality, we consider the wide-baseline reconstruction task where the cameras are in a non-parallel configuration. We define the rotation and translation between the cameras using the matrix $\mathbf{R}$ and vector $\mathbf{t}$ respectively. It follows that plane i induces a homography

$$\mathbf{H}_i = \mathbf{R} + \mathbf{t}\boldsymbol{\alpha}_i^\top, \tag{1}$$

between the images from the cameras.

We define $\mathbf{x}^r$ as the homogeneous representation of a pixel position in the reference image I^r which is part of superpixel s_i. It is then possible to obtain the corresponding pixel location $\mathbf{x}^t$ in the target image using the corresponding oriented plane

$$\mathbf{x}^t = \mathbf{K}_t \mathbf{H}_i \mathbf{K}_r^{-1} \mathbf{x}^r, \tag{2}$$

where $\mathbf{K}_r$ and $\mathbf{K}_t$ are the intrinsic calibration matrices of the reference and target cameras respectively. For compactness we define the function $\mathrm{H}\left(\mathbf{x}_j^r | \boldsymbol{\alpha}_i\right)$ to do this transformation, conditioned on the plane parameters.

Given this, it is now easy to formulate a number of standard appearance matching functions in terms of the oriented plane primitives. Note that, in a slight abuse of notation, the following equations index the 2D images directly using 3 element homogeneous pixel locations ($\mathbf{x}$). The conversion to non-homogeneous co-ordinates is omitted for simplicity.

4.1. Appearance matching

The simplest bottom-up cost function for stereo matching is to employ the Brightness Constancy assumption. We define this as

$$E_{bc}\left(s_i\right) = \sum_{\mathbf{x}_i^r \in s_i} \psi\left(\mathrm{I}^r(\mathbf{x}_i^r) - \mathrm{I}^t(\mathrm{H}\left(\mathbf{x}_i^r | \boldsymbol{\alpha}_i\right))\right). \tag{3}$$

where ψ is a robust cost function.

Similarly we can define matching costs based on the Gradient Constancy assumption

$$E_{gc}\left(s_i\right) = \sum_{\mathbf{x}_i^r \in s_i} \psi\left(\boldsymbol{I}_\Delta^r(\mathbf{x}_i^r) - \boldsymbol{I}_\Delta^t(\mathrm{H}\left(\mathbf{x}_i^r | \boldsymbol{\alpha}_i\right))\right), \tag{4}$$

(where $\boldsymbol{I}_\Delta$ is a gradient image) and the Modified Census Transform [38]

$$E_{ce}\left(s_i\right) = \sum_{\mathbf{x}_i^r \in s_i} \psi\left(\boldsymbol{I}_C^r(\mathbf{x}_i^r) \oplus \boldsymbol{I}_C^t(\mathrm{H}\left(\mathbf{x}_i^r | \boldsymbol{\alpha}_i\right))\right), \tag{5}$$

where $\boldsymbol{I}_C$ are the census transform images. The symbol $\oplus$ represents and "exclusive or" operation, used to calculate the Hamming distance between two vectors.

4.2. Triangulation

In addition to these global cost functions, it is possible to integrate local matching and triangulation criteria within the same framework. Although sparse, these local matching costs tend to have very high confidence, which can make them valuable for avoiding local minima during optimisation. The descriptors used for feature point matching are computed with a pre-trained (using correspondences from the Middlebury07 dataset [1]) deep network including 6 convolutional layers, each of which is followed by max-pooling, subsampling and rectification layers [35]. This produces a descriptor ($\omega \in \mathbb{R}^{128}$). The set of correspondences $\mathcal{C}$ between the two images is then defined by the cosine similarity of the descriptors

$$\mathcal{C} = \left\{ \left(\mathbf{x}_i^r, \mathbf{x}_j^t\right) \middle| \frac{\omega_i^r \cdot \omega_j^t}{\| \omega_i^r \| \| \omega_j^t \|} > \lambda \right\}, \tag{6}$$

where

$$\omega_i^r = \mathrm{CNN}\left(\mathrm{I}^r\left(\mathbf{x}_i^r\right)\right) \quad \text{and} \quad \omega_j^t = \mathrm{CNN}\left(\mathrm{I}^t\left(\mathbf{x}_j^t\right)\right). \tag{7}$$

Figure 2: An example of an "Origami world" interpretation of a scene [8, 9]. The colour indicates the orientation of the surface (i.e. the surface normal vector) at every point. The 3 color channels indicate the 3 components of the normal vector. Blue is the x component, green is the y component and red is the z component. Concave edges between surfaces are also displayed (no convex edges were detected in this example).

These correspondences are then triangulated to produce estimated depths. In order to provide robustness to errors in the camera calibration, we calculate the maximum likelihood depth value for the given correspondence [14]

$$\hat{d} = \min_{d} \sum_{\mathbf{x} \in \{\mathbf{x}^r, \mathbf{x}^t\}} |\mathbf{P} d\mathbf{r} - \mathbf{x}|, \text{ where } \{\mathbf{x}^r, \mathbf{x}^t\} \in \mathcal{C}, \quad (8)$$

where $\mathbf{P}$ is the projection matrix for the camera $\mathbf{x}$ belongs to. The residual of this minimisation $\hat{\nu}$ is also maintained as a confidence score for the quality of the triangulation.

We can now introduce a cost function to exploit this information. As mentioned previously, triangulation results become less reliable as depth increases. To account for this, we can penalise inconsistencies in the inverse depth. Just like the constancy based costs (equations 3-5) this provides a confidence based on what is theoretically observable from the image data [23], ensuring a smooth transition between our different information sources.

Inconsistencies in inverse depth (also known as the fractional depth error) can be penalised by $(\hat{d} - d)/d$ where $\hat{d}$ is the fixed depth measurement and d is the refined depth. For a pixel i, with a fixed depth estimate, this can be re-arranged in terms of the corresponding plane parameters $\boldsymbol{\alpha}_i$

$$\frac{\hat{d}_i - d_i}{d_i} = \frac{1}{d_i}\hat{d}_i - 1 = \mathbf{r}_i^\top \boldsymbol{\alpha}_i \hat{d}_i - 1. \quad (9)$$

We can now create a cost function over all the triangulated matches

$$E_{tr}(s_i) = \sum_{d_i \in s_i} \hat{\nu}_i \psi \left(\mathbf{r}_i^\top \boldsymbol{\alpha}_i \hat{d}_i - 1 \right). \quad (10)$$

5. Top down reconstruction

By formalising the previous bottom-up techniques in terms of oriented planar primitives ($\boldsymbol{\alpha}$), we have retained the ability to enforce top-down constraints on the reconstruction. This can help reconstruction, greatly disambiguating between solutions by exploiting knowledge of the properties of real environments.

The first such cue we exploit is reasoning about the surface normals and their relationship to classes of edge in the scene. To enable this, we follow the data-driven approach of Fouhey *et al*. [9]. Common configurations of surface normal and edge class (Concave, convex and occlusion edges) are recognised, and probabilistically extrapolated to create an "Origami world" interpretation of the scene as shown in figure 2. We can use these estimated surface normal maps to generate an additional matching cost between surface normal images $\boldsymbol{I}_s$

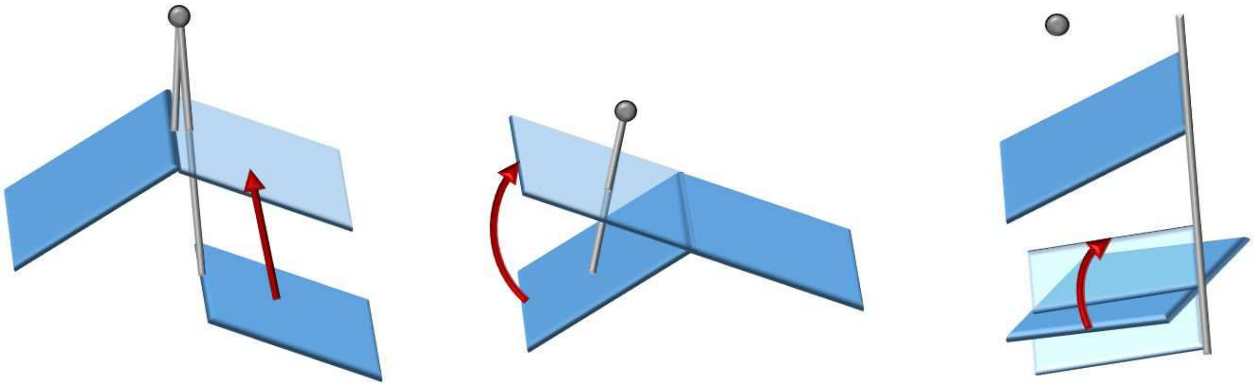
$$E_{sn}(s_i) = \sum_{\mathbf{x}_i^r \in s_i} \psi \left(\mathbf{R} \boldsymbol{I}_s^r(\mathbf{x}_i^r) - \boldsymbol{I}_s^t(\mathrm{H}(\mathbf{x}_i^r | \boldsymbol{\alpha}_i)) \right). \quad (11)$$

Note that the surface normal is rotated by $\mathbf{R}$ before matching, to obtain the expected surface normal in the frame of the other target camera.

We can also introduce top-down pairwise constraints on the relationships between pairs of oriented planes. For example, if two neighbouring superpixels s_i and s_j are not detected as an occlusion boundary, we can favour reconstructions with a concave or convex (rather than disjoint) connection between the surfaces. If we define $\mathcal{N}_{i,j}$ as the set of pixels in s_i which border s_j then the fractional depth error across the boundary corresponds to $\frac{d_j - d_i}{\sqrt{d_i d_j}}$. Which can be re-arranged in terms of $\boldsymbol{\alpha}$ as

$$E_{co}(s_i, s_j) = \sum_{\mathbf{x}_i \in \mathcal{N}_{i,j}} \sum_{\mathbf{x}_j \in \mathcal{N}_{j,i}} \psi \left(\sqrt{d_i d_j} \left(\mathbf{r}_j^\top \boldsymbol{\alpha}_j - \mathbf{r}_i^\top \boldsymbol{\alpha}_i \right) \right). \quad (12)$$

This idea is illustrated graphically in figure 3a. Note that if the superpixels s_i and s_j do not share a boundary (or if it is



(a) E_{co} connects planes in 3D (i.e. encourages boundaries to align), if they are neighbours in the image plane, and are not recognised as an occlusion configuration.

(b) E_{cp} makes connected planes coplanar (i.e. assumes that they are subsections of a larger plane), if they are not recognised as being in a concave or convex configuration.

(c) E_{cl} causes disjoint planes which lie along the same 2D line in the image, to also lie along the same 3D line in the reconstruction.

Figure 3: A visual illustration of the reasoning behind the first 3 pairwise cues.

detected as an occlusion boundary), then the neighbourhood sets $\mathcal{N}_{,}$ are empty.[1]

Similarly if the boundary of two superpixels is not detected as a strongly convex or concave edge, we can integrate a cue which will favour reconstructions integrating them into a larger planar surface. This encompasses the intuition that scenes often contain large planar surfaces, in addition to clutter objects, and can be see as a relaxed Manhattan world assumption (shown in figure 3b). We can enforce this coplanarity constraint by transferring the plane parameters $\boldsymbol{\alpha}_i$ to superpixel s_j (and visa versa) and penalising the fractional depth change which arises over the superpixels

$$\begin{aligned} E_{cp}\left(s_i, s_j\right) = & \gamma_{cp} \sum_{\mathbf{x}_i \in s_i} \psi\left(\sqrt{d_i d_j}\left(\mathbf{r}_i^{\top} \boldsymbol{\alpha}_j - \mathbf{r}_i^{\top} \boldsymbol{\alpha}_i\right)\right) + \\ & \gamma_{cp} \sum_{\mathbf{x}_j \in s_j} \psi\left(\sqrt{d_i d_j}\left(\mathbf{r}_j^{\top} \boldsymbol{\alpha}_j - \mathbf{r}_j^{\top} \boldsymbol{\alpha}_i\right)\right), \end{aligned} \tag{13}$$

where γ_{cp} is an indicator function for non-convex/non-concave edges.

A relaxation of the Manhattan world assumption can be encoded as a collinearity constraint. Intuitively a straight 2D line in the image is likely to arise from a straight 3D line in the environment. Although technically there are an infinite number of 3D curves which would produce a straight 2D projection, most of these curves would be very sensitive to changes in viewpoint. As such, a 2D line is *a priori* much more likely to correspond to a straight 3D line, unless there is strong contrary evidence from other sources of information (see figure 3c).

We can incorporate this idea in a similar way to the coplanarity principle. We define $\mathcal{N}_{i,\bar{\mathbf{e}}}$ as the set of pixels on the border of superpixel s_i and the 2D line $\bar{\mathbf{e}}$. The error is then computed as

$$\begin{aligned} E_{cl}\left(s_i, s_j\right) = & \sum_{\mathbf{x}_i \in \mathcal{N}_{i,\bar{\mathbf{e}}}} \psi\left(\sqrt{d_i d_j}\left(\mathbf{r}_i^{\top} \boldsymbol{\alpha}_j - \mathbf{r}_i^{\top} \boldsymbol{\alpha}_i\right)\right) + \\ & \sum_{\mathbf{x}_j \in \mathcal{N}_{j,\bar{\mathbf{e}}}} \psi\left(\sqrt{d_i d_j}\left(\mathbf{r}_j^{\top} \boldsymbol{\alpha}_j - \mathbf{r}_j^{\top} \boldsymbol{\alpha}_i\right)\right). \end{aligned} \tag{14}$$

Again note that when the superpixel does not border the edge $\bar{\mathbf{e}}$, the corresponding neighbourhood is empty and the cost function is 0.

The final top-down constraint we exploit is to enforce the convexity/concavity of edges between neighbouring superpixels, for recognised configurations. If ϕ_i is the angle between ray $\mathbf{r}_i$ and the ray intersecting the superpixel boundary, then the concavity/convexity is indicated by $\sin(\phi_i)(\hat{d}_j - d_i)$ as shown in figure 4. We can then build a cost function

$$\begin{aligned} E_{ed}\left(s_i, s_j\right) = & \sum_{\mathbf{x}_i^r \in s_i} \psi_{ed}\left(\sin(\phi_i)\left(\mathbf{r}_i^{\top} \boldsymbol{\alpha}_j - \mathbf{r}_i^{\top} \boldsymbol{\alpha}_i\right)\right) + \\ & \sum_{\mathbf{x}_j^r \in s_j} \psi_{ed}\left(\sin(\phi_j)\left(\mathbf{r}_j^{\top} \boldsymbol{\alpha}_i - \mathbf{r}_j^{\top} \boldsymbol{\alpha}_j\right)\right). \end{aligned} \tag{15}$$

Note that the scoring function (ψ_{ed}) applied is different to the other cues. In this case a linear mapping is applied, based on the estimated concave/convex edge probabilities.

6. Optimisation

We combine these unary and pairwise cues over the plane primitives, into a single cost function

$$\begin{aligned} E = & \sum_{s_i \in \mathcal{S}} E_{bc}(s_i) + E_{gc}(s_i) + E_{ce}(s_i) + E_{tr}(s_i) + E_{sn}(s_i) + \\ & \sum_{s_i \in \mathcal{S}} \sum_{s_j \in \mathcal{S}} E_{co}(s_i, s_j) + E_{cp}(s_i, s_j) + E_{cl}(s_i, s_j) + E_{ed}(s_i, s_j). \end{aligned} \tag{16}$$

[1] We define the sum over an empty set to be zero

Technique	Avg. Err.	RMS Err.	A99	Time
BSM	23.5/ 5	52.2/ 5	204/ 5	196/ 5
SGBM1	16.1/ 3	41.5/ 4	180/ 4	0.18/ 1
SGM	8.51/ 1	22.7/ 1	106/ 2	0.99/ 3
SGBM2	16.2/ 4	40.9/ 3	177/ 3	0.29/ 2
Top-down (us)	141/ 5	153/ 5	261/ 5	1.2/ 4
Bottom-up (us)	14.6/ 2	26.1/ 2	123/ 1	3.2/ 4
Full HLSC (us)	13.2/ 2	24.0/ 2	98.1/ 1	3.7/ 4

Technique	Avg. Err.	RMS Err.	A99	Time
LAMC_DSM	14.6/ 7	38.4/ 8	172/ 8	520/ 9
Cens5	10.6/ 4	27.0/ 5	120/ 6	1.34/ 4
SGM	7.63/ 1	21.2/ 1	98.5/ 2	6.48/ 6
SNCC	10.4/ 3	26.3/ 4	110/ 4	0.97/ 3
LPS	12.8/ 5	30.0/ 6	124/ 7	9.35/ 7
IDR	8.57/ 2	23.8/ 2	107/ 3	0.34/ 1
ELAS	15.3/ 8	31.1/ 7	116/ 5	0.72/ 2
SGBM1	16.2/ 9	42.0/ 9	183/ 9	1.48/ 5
HLSC (us)	12.9/ 6	24.0/ 3	91.1/ 1	11.7/ 8

Table 1: Comparison of rankings against the top performing techniques on the Middlebury 2014 benchmark. Comparison is performed against all published techniques for the 2 different resolutions of input data, Quarter size (left) and Half size (right). The ranks in the left and right tables are out of 5 and 9 respectively.

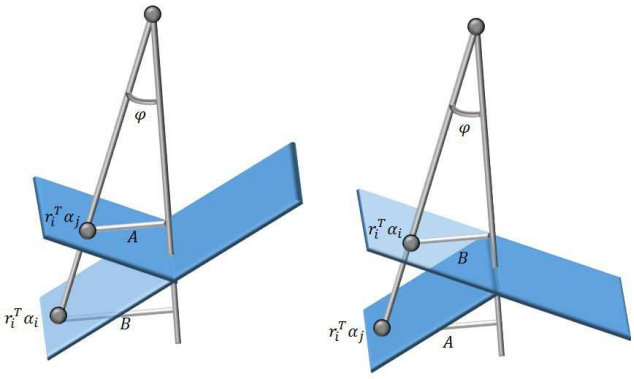


Figure 4: An illustration of the convexity/concavity cost (E_{ed}) from equation 15. Left shows an example of 2 planes (i and j) in a concave configuration; right shows a convex example. A ray intersecting plane i can be projected onto an extrapolation of plane j (transparent). The relative sizes of edges A and B depend on the degree of convexity.

Each energy term has a weighting, which is applied at the same time as the robust scoring function (ψ). These weightings are collectively referred to as $\boldsymbol{\upsilon}$.

The conditional likelihood of the plane parameters is

$$P(\alpha|\mathrm{I}^r, \mathrm{I}^t, \boldsymbol{\upsilon}) = \exp(-E). \tag{17}$$

The optimal values for the weightings $\boldsymbol{\upsilon}$ may then be approximately learned from example data, using Multi-Conditional Learning [30]. This technique approximates the graphical representation of the system as a set of marginal conditional likelihoods.

We then perform MAP inference, maximizing the conditional likelihood to estimate α. Note that all the cost functions are linear in terms of α, with the exception of the image lookups in section 4.1. We therefore linearise these image lookups using a first order Taylor expansion (see supplementary material for more details). This is similar to the approach used to derive the "optical flow constraint" in the motion estimation literature. We are then able to perform efficient inference by solving a Linear Program, while exploiting the high degree of sparsity.

7. Evaluation

We evaluate the proposed approach on the recent Middlebury 2014 dataset [32]. The dataset consists of 33 pairs of high definition (≈6 megapixel) stereo images. For timings, the system was implemented in Matlab and run on a single core at 2.4 GHz. For a full breakdown of performance against image scale (and additional results on the KITTI driving dataset), see the supplementary material. The supplementary material also contains additional experiments on the effect of stereo-baseline and the robustness of monocular cues to viewpoint change.

Our method has very few parameters. The threshold λ for CNN matching during triangulation was set to 0.5. The ground truth for the older Middlebury 2006 [20] dataset was used to learn the optimal weightings $\boldsymbol{\upsilon}$. In addition we found that the best performing cost function (ψ) was the L2 norm. For the superpixel segmentation, we used the efficient graph-based approach of Felzenszwalb and Huttenlocher [7], with a default segmentation threshold of 40.

Two examples of reconstruction for Full resolution (6 megapixel) inputs are shown in figure 5. A number of additional half resolution examples are also shown in figure 6. In table 1 we display the overall results of the Middlebury evaluation for the Quarter resolution and Half resolution benchmarks. We compare against the other currently published techniques which evaluate on each resolution. The performance is computed for fully dense estimates, including occluded regions. We tabulate the the average and RMS error in terms of disparity levels to give an idea of overall accuracy. In addition we tabulate the 99th percentile error (referred to as A99 in the Middlebury2014 benchmark), which provides an indication of the quantity and magnitude of outliers in the reconstruction. This can be seen as a measure of robustness (i.e. catastrophically incorrect interpretations of

Technique	Avg. Err.	RMS Err.	A99	Time
SGM	4.90 / 2	16.2 / 2	86.8 / 2	55.4 / 5
PFS	4.83 / 1	17.2 / 3	97.4 / 4	28.4 / 3
ELAS	6.28 / 4	18.9 / 4	94.5 / 3	4.12 / 1
LPS	7.63 / 5	25.9 / 5	143 / 5	29.3 / 4
SGMB1	21.1 / 6	48.2 / 6	177 / 6	13.9 / 2
HLSC (us)	5.91 / 3	13.1 / 1	66.5 / 1	99.8 / 6

Technique	Avg. Err.	RMS Err.	A99	Time
SGM	4.65 / 1	14.7 / 1	79.0 / 2	13.3 / 5
PFS	6.89 / 2	20.9 / 4	109 / 5	5.55 / 3
ELAS	7.72 / 3	19.3 / 3	83.0 / 3	0.89 / 1
LPS	16.4 / 5	31.6 / 5	98.0 / 4	8.39 / 4
SGMB1	17.8 / 6	43.8 / 6	193 / 6	2.93 / 2
HLSC (us)	8.04 / 4	17.6 / 2	78.0 / 1	45.3 / 6

Table 2: Detailed comparison of 2 sequences (*Adirondack* left, *ArtL* right) in the Full resolution benchmark. Ranks out of 6.

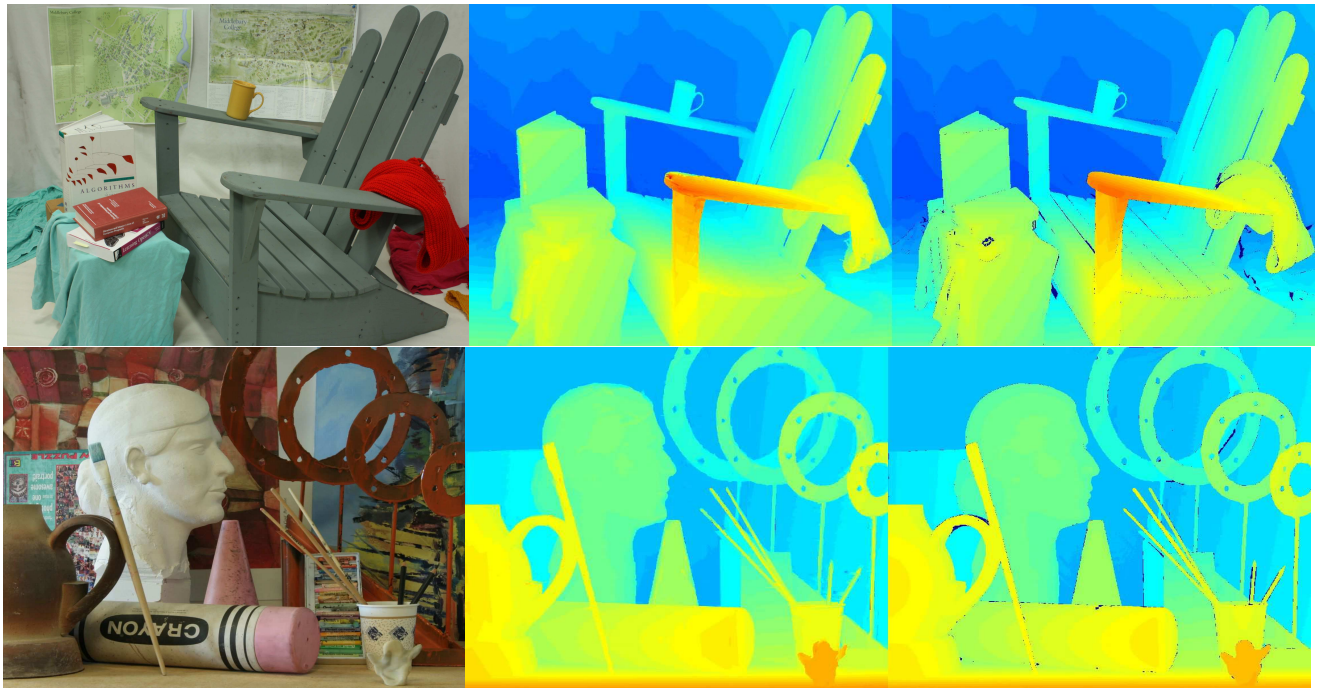


Figure 5: Example Full resolution reconstructions from the Middlebury 2014 dataset. One input image (left), the output of our algorithm (middle) and the ground truth (right).

the scene). Lower is better for all performance measures.

Additionally we contrast the proposed technique using different types of cue. Using only top-down reasoning with no matching is significantly faster, however the quality of the estimate is poor as the finer details of the model are no longer refined. It is interesting to note that the decrease in robustness (roughly a factor of 3) is significantly lower than the loss of accuracy (roughly a factor of 10). When the technique exploits only bottom-up matching cues the reconstruction is of higher quality. However, the combination of bottom-up and top-down performs the best, with around 10% improvement in all error measures, reinforcing the complementary nature of the different cues, particularly improving robustness by resolving ambiguities.

For the full resolution benchmark, we examine in detail the performance of the relevant techniques for the *Adirondack* and *ArtL* sequences in table 2. At these higher resolutions our technique remains the most robust algorithm. However, accuracy is drastically improved. The increase in resolution makes it possible to use smaller planar-primitives, without the optimisation problem becoming ill conditioned. These smaller primitives make it possible to model fine scene details, with improved fidelity.

In table 3 we examine the performance for each of the 15 training sequences in the Half resolution benchmark, where ground truth is provided (additional results are included in the supplementary material). We list the performance in each of the 3 categories, along with the ranking out of 25 (including unpublished techniques, and techniques running on different resolution data). As previously noted, the high level scene cues make the algorithm extremely robust. This helps reduce outliers in areas of low texture information.

It is interesting to note that the dataset contains 2 scenes (*ArtL* and *PianoL*) with lighting changes between the views and one scene (*MotorcycleE*) with significantly different exposure levels between the views (The suffix P indicates "perfect" calibration, and has very little effect on our algorithm due to reduced reliance on triangulation). Our algorithm is extremely resilient to these changes in lighting and exposure compared to traditional bottom-up reconstruc-

Figure 6: Example Half resolution reconstructions from the Middlebury 2014 dataset. Each triplet shows one input image (left), the output of our algorithm (middle) and the ground truth (right).

	Adirondack	ArtL	Jadeplant	Motorcycle	MotorcycleE	Piano	PianoL	Pipes
Avg. Err.	5.76 / 15	9.91 / 15	39.6 / 14	9.85 / 17	9.54 / 13	8.09 / 17	10.8 / 8	15.0 / 20
RMS Err.	13.3 / 8	21.3 / 15	75.4 / 5	26.2 / 17	25.9 / 13	12.2 / 9	15.3 / 3	36.7 / 19
A99	63.3 / 5	75.2 / 4	283 / 2	140 / 17	139 / 13	43.6 / 6	56.2 / 2	149 / 14
	Playroom	Playtable	PlaytableP	Recycle	Shelves	Teddy	Vintage	**Average**
Avg. Err.	11.0 / 17	25.2 / 16	14.0 / 22	6.74 / 22	13.3 / 18	2.97 / 11	19.5 / 22	12.9 / 14
RMS Err.	18.6 / 6	22.4 / 3	19.5 / 16	11.6 / 11	20.0 / 11	8.66 / 10	22.7 / 5	24.0 / 10
A99	99.9 / 7	77.9 / 3	59.5 / 13	43.0 / 5	79.2 / 8	43.5 / 8	61.7 / 1	91.1 / 2

Table 3: Analysis of the performance on the 15 training sequences where ground truth is provided on the Half resolution benchmark. Listed is the error value for that sequence, followed by the ranking out of 25 for that sequence.

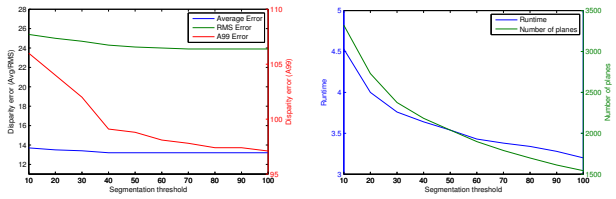


Figure 7: Behaviour of the approach with different segmentation thresholds. Left - Plots of the 3 accuracy characteristics. Right - plots of the tradeoff (speed and number of planes). Note that both subfigures display two Y scales.

tion techniques. Performance on the *Motorcycle* sequence with and without the exposure change are roughly the same, which leads to a 25% improvement in ranking, as other techniques are adversely affected. For the lighting change in the Piano sequence, performance drops by around 20%, however this is dramatically lower than most other techniques, leading to an increase in ranking of over 60%.

We also evaluate the effect of varying the superpixel segmentation threshold in figure 7 using the Quarter Resolution benchmark. Higher thresholds lead to a smaller numbers of larger superpixels, and can significantly improve the runtime of the algorithm. However, the effect on accuracy is negligible for thresholds of 40 and over. Below 40, the superpixels are often poorly constrained due to their small size, and accuracy suffers.

8. Conclusions

From these results we can conclude that, as in human vision, automatic reconstruction benefits greatly from top-down reasoning about the environment. Furthermore, the proposed fusion framework using slanted plane primitives has proven a powerful and highly efficient approach to achieving this. We have demonstrated also the flexibility of this approach, incorporating a vast array of different information sources within a single unified scheme.

In the future, the automatic learning of cue weights (Section 6) could be extended to recognise particular types of environment, and either use weightings specialised to that type of environment, or to even perform online estimation of the cue weights for temporal stereo. It would also be beneficial to explore ways to integrate "recognition meets reconstruction" into the framework. This is an extremely powerful top down cue, and could be extended further by including relationships between recognised entities. In addition it may prove interesting to extend the CNN component to estimating larger parts of the representation, rather than only the triangulation cues.

Acknowledgements

This work was supported by the EPSRC project Learning to Recognise Dynamic Visual Content from Broadcast Footage (EP/I011811/1). The authors also wish to thank David Fouhey for providing code and advice relating to the estimation of monocular cues.

References

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011.

[2] R. M. Batson. Photogrammetry with surface-based images. *Applied optics*, 8(7):1315–1322, 1969.

[3] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. Pmbp: Patchmatch belief propagation for correspondence field estimation. *IJCV*, 110(1):2–13, 2014.

[4] M. Bleyer, C. Rhemann, and C. Rother. Extracting 3d scene-consistent object proposals and depth from stereo images. volume 7576, pages 467–481. Springer Berlin Heidelberg.

[5] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereojoint stereo matching and object segmentation. pages 3081–3088, June.

[6] J. E. Cutting. *Perception of space and motion*, chapter The Integration, Relative Potency, and Contextual Use of Different Information about Depth, page 69. 1995.

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

[8] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3D primitives for single image understanding. In *ICCV*, 2013.

[9] D. F. Fouhey, A. Gupta, and M. Hebert. Unfolding an indoor origami world. In *ECCV*, 2014.

[10] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. pages 25–38.

[11] F. Guney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. June.

[12] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. volume 6314, pages 482–496.

[13] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation.

[14] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University press, 2000.

[15] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. pages 224–237.

[16] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV '09)*, 2009.

[17] P. Heise, S. Klose, B. Jensen, and A. Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2360–2367, 2013.

[18] S. Hermann and R. Klette. Iterative semi-global matching for robust driver assistance systems. pages 465–478.

[19] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[20] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, pages 1–8. IEEE, 2007.

[21] M. Humenberger, T. Engelke, and W. Kubinger. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In *CVPR Workshop*, 2010.

[22] B. Julesz. Binocular depth perception of computer-generated patterns. *Bell System Technical Journal*, 39(5):1125–1162, 1960.

[23] R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. pages 55–71.

[24] A. Kumar, S. Haker, C. Vogel, A. Tannenbaum, and S. Zucker. Stereo disparity and l1 minimization. In *Conference on Decision and Control*, 1997.

[25] G. Kuschk and D. Cremers. Fast and accurate large-scale stereo reconstruction using variational methods. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 700–707, Dec 2013.

[26] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized RANSAC. pages 1013–1023, London, UK, September 2012. BMVA.

[27] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. pages 2136–2143. IEEE.

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91 – 110, November 2004.

[29] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu. Constant time weighted median filtering for stereo matching and beyond. pages 49–56, Dec.

[30] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Conference on Artificial Intelligence*, volume 21, page 433, 2006.

[31] R. Ranftl, T. Pock, and H. Bischof. Minimizing tgv-based variational models with non-convex data terms. In A. Kuijper, K. Bredies, T. Pock, and H. Bischof, editors, *Scale Space and Variational Methods in Computer Vision*, volume 7893, pages 282–293. Springer Berlin Heidelberg, 2013.

[32] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42, 2014.

[33] R. Spangenberg, T. Langner, and R. Rojas. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *Computer Analysis of Images and Patterns*, pages 34–41, 2013.

[34] X. Wang, D. F. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation.

[35] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching.

[36] K. Wu and M. Levine. Recovering parametric geons from multiview range data. pages 159–166, Jun.

[37] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. pages 1862–1869, June.

[38] R. Zabih and J. Woodfill. A non-parametric approach to visual correspondence. *PAMI*, 1996.

[39] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition*, pages 214–223. Springer, 2007.

[40] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-rigid-as-possible stereo under second order smoothness priors.